The Dissertation Committee for Bum Kyu Lee certifies that this is the approved version of the following dissertation:

Genome-wide Target Identification of Sequence-specific Transcription Factors through ChIP Sequencing

Committee:

Vishwanath Iyer, Supervisor

Arturo DeLozanne

Edward Marcotte

Philip Tucker

Scott Stevens

Genome-wide Target Identification of Sequence-specific Transcription Factors through ChIP Sequencing

by

Bum Kyu Lee, B.E.; M.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2011

Dedication

I dedicate this work to my parents who support me to achieve my goals

Acknowledgements

I would like to thank my advisor, Dr. Vishwanath Iyer for his patience, scientific guidance and support throughout graduate school. Without his constant encouragement, guidance and mentorship I would not have achieved my goals.

I would also like to thank my committee members Dr. Arturo DeLozanne, Dr. Edward Marcotte, Dr Philip Tucker, and Dr. Scott Stevens for thoughtful advices and suggestions which guide me to perform my projects.

I want to thank all Iyer lab members: Daechan Park, Yaelim Lee, Damon Polioudakis, Dia Bagchi, Yunyun Ni for discussion on my projects, especially Dr. Akshay Binge who helped me learn basic computational skills as well as gave me invaluable advices for data analysis and Anna Battenhouse who helped me not only analyze large scale data but also generate many figures for papers as well as editing my dissertation.

I also want to thank all my friends, Hyung-Chul Kim, Dae-Suk Eeom, ji-Hoon Lee, Young-Sam Lee, Yong-Hwan Kim, and Hae-Ryung Chang who helped me get through core courses and gave invaluable advices to prepare my Part 1 exam.

Finally, I would like to thank my parents who have been providing me ceaseless support and encouraging me to pursue my graduate studies without fail.

Genome-wide Target Identification of Sequence-specific Transcription Factors through ChIP Sequencing

Publication No._____

Bum Kyu Lee, Ph.D.

The University of Texas at Austin, 2011

Supervisor: Vishwanath R. Iyer

The regulation of gene expression at the right time, place, and degree is crucial for many cellular processes such as proliferation and development. In addition, in order to maintain cellular life, cells must rapidly and appropriately respond to various environmental stimuli. Sequence-specific transcription factors (TFs) can recognize functional regulatory DNA elements in a sequence-specific manner so that they can regulate only a specific group of genes, a process which enables cells to cope with diverse internal and external stimuli. Human has approximately 1,400 sequence-specific TFs whose aberrant expression causes a wide range of detrimental consequences including developmental disorders, diseases, and cancers; therefore, it is pivotal to identify the binding sites of each sequence-specific TF in order to unravel its roles in and mechanisms of gene regulation.

Even though some TFs have been intensively studied, the majority of TFs still remain to be studied, particularly the tasks of identifying their genome-wide target genes and deciphering their biological roles in specific cellular contexts. Many questions remain unanswered: how many sites on the human genome a sequence-specific TF can bind; whether all TF-bound sites are functional; how a TF achieves binding specificity onto its targets; how and to what extent a TF is involved in gene regulation. Comprehensive identification of the binding sites of sequence-specific TFs and follow-up molecular studies including gene expression microarrays will provide close answers to these questions.

Chromatin Immunoprecipitation coupled with recently developed high-throughput sequencing (ChIP-seq) allows us to perform genome-scale unbiased identification of the binding sites of sequence-specific TFs. Here, to gain insight into gene regulatory functions of TFs as well as their influences on gene expression, we conducted, in diverse cell lines, genome-wide identification of the binding sites of several sequence-specific TFs (CTCF, E2F4, MYC, Pol II) that are involved in a wide range of biological functions, including cell proliferation, development, apoptosis, genome stability, and DNA repair. Analysis of ChIP-seq data provided not only comprehensive binding profiles of those TF across the genome in diverse cell lines, but also revealed tissue-specific binding of CTCF, MYC, and Pol II as well as combinatorial usage among these three factors. Analyses also showed that some CTCF binding sites were inherited from parents to children and regulated in an individual-specific as well as allele-specific manner.

Finally, genome-wide target identification of several TFs will broaden our understanding of the gene regulatory roles of these sequence-specific TFs.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 COMPLEXITY AND DYNAMICS OF THE GENOME IN EUKARYOTES

Unlike prokaryotic genomes, eukaryotic genomes are packaged in vivo into compact DNA-protein complexes known as chromatin in the nucleus. Chromatin is composed of nucleosomes where 147 base pairs of DNA are wrapped 1.65 times around a core histone octamer (two copies of each H2A, H2B, H3, and H4) in a left-handed toroid (Luger et al, 1997). Nucleosomes can form a linear array along with DNA, and can further be compacted by H1 into higher-order 30nm fibers that are transcriptionally inactive (Campos & Reinberg, 2009). Thanks to this compact and highly-ordered chromatin structure, cells are able to carry an enormous amount of genetic information in the nucleus. For instance, the human genome has about 25,000 protein coding genes (J. Craig et al, 2001); the fruit fly *Drosophila melanogaster* ~13,000 (Adams et al, 2000); the rockcress plant *Arabidopsis thaliana* ~25,000 (The Arabidopsis Genome Initiative, 2000); and the nematode worm *Caenorhabditis elegans* ~20,000 (C. elegans Sequencing Consortium, 1998). Chromatin structure is well conserved from yeast to humans, although mammalian chromatin has additional complexity required for cell differentiation (Rando & Chang, 2009).

In addition to this complex chromatin structure, eukaryotic genomes are temporally and spatially organized in the nucleus. For instance, it is widely accepted that

gene-rich regions in interphase chromosomes tend to localize in the nuclear interior whereas gene-depleted regions occupy more peripheral parts of the nucleus (Federico et al, 2006). This complex and dynamic organization of the genome emerges as a key regulatory determinant of gene expression by establishing and maintaining active, poised, or repressive chromatin states (Misteli, 2007; Schneider & Grosschedl, 2007).

The interphase genomes in eukaryotes can be broadly classified into euchromatin and heterochromatin domains. Euchromatin replicates early and locates in the central position of the nucleus, whereas heterochromatin replicates late and occupies more peripherally in the nucleus (Wood et al, 2010) . In addition, euchromatin is not only less condensed but also actively transcribed while heterochromatin is tightly packed as well as transcriptionally inert (Manuelidis, 1991). Furthermore, recent study has revealed the link between chromatin structure and gene density, but not the status of gene activity, implying that more complex regulatory mechanisms are involved in gene expression such as DNA or posttranscriptional histone modification (Gilbert et al, 2004; Spector, 2004).

Overall, eukaryotic genomes exist with the compact chromatin structure which provides poor DNA access to TFs and other DNA binding proteins. In order to maintain cellular homeostasis and properly respond to various internal and external stimuli as well as to accomplish cell proliferation, development, and DNA replication and repair, chromatin structures must be locally and dynamically adjusted into open structures either by de-condensing chromatin or looping out DNA (Campos & Reinberg, 2009).

Two types of mechanisms exist to affect the temporal changes in chromatin structures needed to provide access to the DNA for transcriptional machineries. The first is multiple covalent posttranslational modification, which epigenetically modulates gene expression by changing histone tails in eight distinctive manners: acetylation, methylation, phosphorylation, ubiquitylation, sumoylation, ADP-ribosylation, deimination, and proline-isomerization (Kouzarides, 2007). These modifications are implicated in diverse biological processes including DNA condensation, transcription, and DNA replication as well as repair. A wide variety of combinatory histone modifications have been identified, and these modifications can either promote or impede the access of transcriptional machineries by altering intrinsic properties of chromatin, in particular neutralizing their positive charge (Hansen, 2002).

Based on the presence of specific histone-modification marks, chromatin can be classified into active and inactive domains. Recent genome-wide mapping studies of histone modifications using ChIP-seq have revealed that transcriptionally active chromatin is generally enriched with H3K4me1 (in enhancer regions), H3K4me2, H3K4me3 (particularly at 5' ends of gene), H3K36me3 (in gene bodies and 3' end of genes), H3K9me1, H3K27me1, and H4K20me1; whereas repressive chromatin exhibits increased hypo-acetylation, H3K9 methylation, and H3K27me3 (Barski et al, 2007); (Wang et al, 2008). Interestingly, it has also been reported that repressive marks in active chromatin (and vice versa) as well as "bivalent chromatin domains" with both active (H3K4me3) and repressive (H3K27me3) histone marks, can coexist, particularly in

development-related genes of ES cells (Bernstein et al, 2006). These complicated histone codes contribute to more sophisticated transcriptional gene regulation.

The second mechanism of chromatin structure transformation is a chromatin modifying complex, or chromatin remodeler that utilizes ATP to remove, destabilize, or reorganize nucleosomes. Several chromatin-remodeling complexes including SWI/SNF, SWR, ISW, CHD, NuRD, and INO80 have been reported in eukaryotes (Kouzarides, 2007). Additionally, advances have recently been made in understanding remodeling functions and mechanisms in different environmental conditions such as heat shock or DNA damage (Larsen et al, 2010; Shivaswamy & Iyer, 2008; Smeenk et al, 2010). However, the roles of many chromatin-remodeling complexes remain unclear, in particular the functions of each component in a complex and the functional interplay between a complex and histone modifications.

In addition to the roles of post-translational histone modification and chromatin remodeling complexes, three-dimensional models of chromatin emerge as an additional sophisticated key regulatory mechanism of gene regulation. Most recently two independent studies have revealed inter-chromosomal as well as intra-chromosomal interactions of DNA in chromatin (Fullwood et al, 2009; Lieberman-Aiden et al, 2009).

It has become clear that rather than simply being a passive DNA packing place, chromatin is a dynamic and active regulator for gene expression in eukaryotes (Lemon & Tjian, 2000). Nucleosome loss or relocation either by histone modifications or chromatin modifying complexes is necessary for proper level of gene expression as well as rapid

transcriptional activation in response to environmental stimuli. However, neither nucleosome removal nor relocation alone is sufficient to fully activate genes. Eukaryotes have several other regulatory machineries for gene expression, including transcription factors. Furthermore, these regulators must orchestrate their effects in order to achieve appropriate, context-specific gene expression (Lemon & Tjian, 2000).

1.2 TRANSCRIPTIONAL REGULATION OF GENE EXPRESSION IN EUKARYOTE

Transcription is a primary step in gene expression. Two main regulatory components for transcription in eukaryotes exit: cis-acting and trans-acting elements. The interaction between these two elements triggers transcription (Maston et al, 2006). Following signal transduction cascades which activate them, transcription factors bind the promoters (cis-acting elements) of genes and recruit co-activators, including histone modifying enzymes and chromatin remodelers. These sequential recruitments further facilitate open chromatin, which in turn facilitates assembly general transcription machinery near the transcription start sites (TSSs) of genes, forming a transcription-initiation complex.

Cis-acting regulators contain proximal regulatory elements, promoters as well as position-independent distal regulatory elements such as enhancers, repressors, silencers, and insulators that are located far away from TSSs (Maston et al, 2006). The general function of promoters is to define the orientation of transcription and provide an area

where basal transcriptional machinery assembles (Maston et al, 2006), and core promoters are highly conserved in orthologous genes (Smale & Kadonaga, 2003). A classical promoter consists of common elements including the TSS, TATA box, and initiator. The TATA box, an AT-rich site generally found 25 to 30bp upstream of the TSSs in metazoa, provides the binding sites for the TATA-binding protein (TBP) (Smale & Kadonaga, 2003). However, TBP can interact with a broad range of sequences instead of the defined canonical one, and several studies have revealed that many promoters are TATA-less and that approximately 50% of human genes have alternative promoters, spreading their regulatory elements over wider distances (Kimura et al, 2006). Furthermore, in metazoa, depending on the genes, promoters can have either only an initiator element or a TATA box, and in some case neither (Lee & Young, 2000; Maston et al, 2006). In addition to the TATA box, a subset of promoters have CpG islands, downstream promoter element (DPE), and TFIIB recognition element (BRD) (Sandelin et al, 2007), all of which can contribute to the recruitment and assembly of transcription machinery.

Until now, tens of thousands promoters in human have been identified through both computational and experimental approaches, including high-throughput full length c-DNA sequencing as well as sequencing-based methods such as rapid amplification of 5’ complimentary DNA ends (5’RACE) (CSH, 2005), cap analysis of gene expression (CAGE) (Shiraki et al, 2003), serial analysis of gene expression (5’-SAGE) (Hashimoto et al, 2004), and paired-end tags (PET) (Ng et al, 2005).

It is well established that many genes are under the control of distal cis-regulatory elements. Numerous types of studies have attempted to identify cis-regulatory elements genome-wide and to elucidate their biological functions. For instance, serial deletion analysis identified several discrete enhancers that functions in a position- and orientation-independent manner and are normally located in far away from the TSSs of genes that they up-regulate; these enhancers assist in the recruitment of TFs onto promoters and can alter the activity of TFs (Valenzuela & Kamakaka, 2006). In vertebrates, enhancers are considered to be an aggregate location of TF binding sites (Panne, 2008) and many genes can be regulated by multiple arrays of enhancers (Visel et al, 2007). Several enhancers for developmental genes were also identified through transgenic mouse assays coupled with LacZ reporter gene assays using developing mouse embryos (Visel et al, 2009). Moreover, recent genome-wide discovery of enhancers, based on histone marks and p300 binding sites obtained from ChIP-seq, make it possible not only to predict many long-range cis-acting regulatory elements but also to reveal some of their tissue-specific functions (Heintzman et al, 2009; Heintzman et al, 2007). Many disease phenotypes are attributed to the genomic rearrangement of enhancers that disrupts the regulation of genes, even for enhancers located remote from the gene targets (Kleinjan et al, 2001).

In contrast to gene activation by enhancers, other elements such as silencers and repressors can suppress gene expression regardless of their position and orientation. The repression caused by silencers is due to the interaction of histones with silencing proteins

such as histone deacetylases. Insulators are not only able to block the communication between promoters and enhancers of genes, but also inhibit the spread of repressive chromatin structures (Valenzuela & Kamakaka, 2006). Many insulators bind to the promoters of genes as well as to the binding sites of TFs, thus helping maintain the active and inactive regulation of gene expression by functionally insulating genes from neighboring positive or negative regulatory elements (Sproul et al, 2005; West et al, 2002). Further studies have revealed the versatile roles of these non-coding cis-regulatory elements which are implicated in establishing various biological functions and lineage-specific roles of diverse tissues (De Lucia & Dean, 2010; Huarte et al, 2010; Orom & Shiekhattar, 2011; Pauli et al, 2011).

Trans-regulatory elements consist of a mediator complex as well as both general and sequence-specific TFs. The mediator complex, composed of multiple proteins, is a general regulator of transcription in eukaryotes and its function is evolutionally conserved from yeast to mammals (Myers & Kornberg, 2000). Recent research has revealed that the mediator complex can link distal cis-regulatory elements (in particular enhancers) to promoters by recruiting a cohesion-containing protein complex which forms a bridge between enhancers and promoters (Kagey et al, 2010). The interaction of mediator complex and distal cis-regulatory elements with promoters further stabilizes the assembly of the transcription pre-initiation complex (Taatjes, 2010). In addition to the mediator complex, general TFs including TATA binding protein (TBP) recognize core promoters where several key functional elements such as TATA boxes and initiators lie,

and form a transcriptional pre-initiation complex which facilitates recruitment of RNA polymerase II (Pol II).

Among many transacting factors, sequence-specific TFs are key cellular components regulating the expression of only a subset of target genes, enabling cells to cope with diverse environmental stimuli (Farnham, 2009). Due to their specificity to target genes, TFs are involved in diverse biological processes including cell proliferation, development, and swift responses to intracellular as well as external stimuli.

1.3 SEQUENCE-SPECIFIC TRANSCRIPTION FACTORS

A sequence-specific TF can be defined as a protein containing both a DNA-binding domain that recognizes DNA in a sequence-specific manner and a trans-acting domain that modulates its downstream target genes. Computational analysis of DNA sequences, based on distinctive properties of TFs, reveal that approximately 2,000 TFs exist in humans, of which about 1,400 have sequence-specific DNA-binding characteristics (Vaquerizas et al, 2009).

Expression analysis of 873 sequence-specific TFs using Affymetrix GeneChips across 32 human samples has revealed that genes of sequence-specific TFs have lower expression than non-TF genes. The low expression level of a TF makes it easier to rapidly alter its concentration (hence its activity), and also promotes precise target discrimination by favoring binding sites with the strongest TF affinity (Vaquerizas et al,

2009). Finally, both target specificity and regulatory flexibility are enabled by combinatory functioning of TFs with other activators. Multiple TFs can bind onto an enhancer element, which is called the enhanceosome (Merika & Thanos, 2001). This combinatorial usage allows very low concentrations of TFs to influence transcription, while the cooperative assembly of enhanceosome variants can provide tissue-specific programming (Levine & Tjian, 2003; Panne et al, 2007).

The most important role of sequence-specific TFs is that they are key players in transcriptional gene regulation, especially in dealing with internal or external demands to maintain cellular life. For example, a heat shock transcription factor (HSF) is activated in response to physiological stresses such temperature elevation or chemical exposure that disrupts metabolic processes (Mosser et al, 1990). Moreover, sequence-specific TFs are expressed either in one cell line or in almost all tissues, implying they can trigger transcriptional gene regulation either in a ubiquitous or a cell-type specific manner (Vaquerizas et al, 2009).

Many tissue-specific TFs have been reported: three retina-specific TFs, cone-rod homeobox protein (CRX), neural retinal-specific leucine zipper (NRL), and nuclear receptor subfamily 2, group E, member 3 (NR2E3) (Qian et al, 2005); a muscle specific TF, MEF-2 (formyocyte-enhancer-binding-factor); a liver-specific TF, hepatocyte nuclear factor3 (HNF3 or FOXA) (Kaestner et al, 1998); and a brain-specific TF, Nuclear receptor related 1 protein (NURR1) which is expressed only in the central nervous

system (CNS) and is crucial for the development of dopamine neurons (Law et al, 1992; Perlmann & Wallen-Mackenzie, 2004).

Numerous diseases including cancers as well as developmental disorders can be caused by the abnormal regulation of transcription. For instance, many TFs have oncogenic properties whose deregulation induces a wide range of tumors (Furney et al, 2006; Jimenez-Sanchez et al, 2001); malfunction of TFs implicated in one-third of human developmental disorders (Boyadjiev & Jabs, 2000); and it has been reported that 164 TFs are directly responsible for 277 diseases (Vaquerizas et al, 2009). Moreover, much phenotypic diversity and evolutionary adaptation are attributed to the adjustment of the activity and regulatory specificity of TFs (Bustamante et al, 2005; Lopez-Bigas et al, 2008).

Until now, only a limited number of human sequence-specific TFs have been well characterized. To better understand the functions of sequence-specific TFs and in particular to increase insights into their physiological roles in different tissues it is pivotal to identify their binding sites across the genome in diverse cell lines under various environmental conditions.

1.4 GENOME-SCALE IDENTIFICATION OF SEQUENCE-SPECIFIC TF BINDING SITES.

Genome-wide mapping of TF binding sites is essential not only for understanding the mechanisms of transcriptional gene regulation, but also for deciphering the gene

regulatory networks implicated in various biological processes. The ENCODE Project Consortium commenced in 2004 with the goal of advancing our understanding of human genome functions (Birney et al, 2004). One of the major tasks of the ENCODE project is to identify and catalogue all possible functional DNA elements in the human genomes using high-throughput experimental methods as well as computational approaches. By investigating 1% of the human genome (30Mb) covering 44 genomic regions including protein-coding and non-coding loci, many novel non-coding transcripts and regulatory sites have been identified (Birney et al, 2007). Furthermore, the ENCODE Consortium is now scaling up its project from 1% to whole genome in order to identify genome-wide functional cis-regulatory elements, including TF binding sites for several dozen TFs in diverse human tissues.

Before ChIP coupled with next generation sequencing method (ChIP-seq) was developed, chromatin immunoprecipitation followed by microarrays (ChIP-chip) was one of the major tools for investigating genome-scale precise locations where protein interacts with DNA in vivo. The conventional ChIP-chip method made it possible not only to identify numerous novel TF binding sites but also to elucidate several unknown functions. For instance, E2F4 and SRF ChIP-chip experiments have revealed several hundred putative binding sites using core-promoter arrays covering -800 to +300 from TSSs in the human genome (Balciunaite et al., 2005, Cooper et al., 2007, Xu et al., 2007). However, the genome-wide identification of TFs binding sites using ChIP-chip focused mainly on proximal promoter regions ranging from 1 kb to 5 kb upstream of known

TSSs; these region-limited studies prevented us from drawing conclusions regarding whether the discovered TFs binding sites were comprehensive across the genome. In addition, conventional ChIP-chip is limited in its ability to pinpoint precise binding sites of TFs due to its low resolution, generally 30-100 bp, even though whole-genome tilling arrays, which produce 5-20bp resolution, partially overcome these drawbacks (Liu, 2007).

More recently, ChIP-seq technology has enabled the genome-scale unbiased identification of TFs binding sites without the scale and resolution limitation of conventional ChIP-chip, and at a cost 3-5 times lower (Park, 2009). In ChIP-seq, the DNA fragments pulled down with a specific antibody are directly sequenced rather than being hybridized on an array. ChIP-seq has several advantages compared with the conventional ChIP-chip: higher resolution (single nucleotide), higher coverage in repetitive regions normally masked out of an array, fewer artifacts caused by cross hybridization, lower requirement for initial ChIP materials, a higher dynamic detection range without the limit of low as well as high signal, and cost effectiveness (Park, 2009).

Several next-generation sequencing methods have been developed in recent years including Solexa (Illumina), Solid (Life Technologies), and 454 (Roche) (Park, 2009). Among them, Solexa and Solid sequencing technology can generate several hundred millions of short (35-50bp) reads in one time sequencing (Metzker, 2010). Thanks to this advance in technology, tremendous progress has been made in identifying protein-DNA interaction sites genome-wide, in particular sequence-specific TFs binding sites and

histone modification sites. For example, studies of histone modification sites have not only established histone marks for active as well as repressive genomic regions but have also defined precise locations of enhancers which in general had strong H3K4me1 occupancy signal but weak H3K4me3 (Kim et al, 2010; Rada-Iglesias et al, 2010

). The comprehensive analysis of TF binding sites using ChIP-seq provides new insights into TFs binding patterns that change depending on environmental conditions, and allows us to understand the consequences of TFs binding onto diverse cis-regulatory element of the genome.

1.5 SUMMARY OF RESEARCH GOALS

In order to broaden our knowledge of the transcriptional regulation of genes and compose a picture of the overall regulatory network of gene expression, genome-wide profiling of each sequence-specific TF's binding sites must be first be obtained. Among ~1,400 sequence-specific TFs, genome-scale target identification has been performed in only a limited number. Here, we investigated the binding sites of several TFs including E2F4, MYC, CTCF and Pol II in diverse cell lines.

In Chapter 2, we investigated genome-wide E2F4 binding sites using ChIP-seq to identify and catalogue all putative binding sites of E2F4 (a member of the E2F family of TFs) across the human genome. In addition, we performed overexpression of E2F4 and its cofactors followed by microarray analysis in order to elucidate the functional

relevance of E2F4 and its cofactors binding to their target gene expression. We also investigated putative E2F4-bound enhancer sites based on published enhancer histone marks, and further validated the enhancer function of some distal E2F4 sites using luciferase reporter gene assays.

In Chapter 3, we examined genome-wide binding sites of CTCF, MYC, and Pol II in diverse cell lines as part of the ENCODE project to characterize the binding preference of each TF for genomic loci having different properties, such as CpG islands and gene-dense regions, and to investigate the influence of each TF binding on its target gene expression. Furthermore, we examined the combinatory usage of these TFs and their influence on gene expression. Finally, we investigated tissue-specific TF binding and its consequences for gene expression.

In Chapter 4, we scrutinized individual-specific and allelic-specific CTCF binding sites in six different individuals including two different parent-daughter trio sets (CEPH and Yoruba families) in order to investigate the influence of single nucleotide polymorphism (SNP) on CTCF recruitment on the genome and subsequent influences of allele-specific binding on gene expression.

Chapter 2: Versatile functions of E2F4 in transcriptional gene regulation revealed by genome-wide analysis

2.1 INTRODUCTION

The E2F transcription factor (TF) family consists of 8 different proteins including E2F1, E2F2, E2F3a, E2F3b (an isoform of E2F3) and E2F4 to E2F8. Those proteins play crucial roles in cell cycle regulation as well as development by activating or suppressing certain classes of E2F responsive genes (Attwooll et al, 2004; Rowland & Bernards, 2006). E2Fs can functions as either activators (E2F1-E2F3) or repressors (E2F4-E2F8) (Crosby & Almasan, 2004). Interestingly, recent research revealed that E2F1-E2F3 can switch from being activators to repressors in differentiating cells (Chong et al, 2009). The expression of E2F1-E2F3 is tightly regulated during cell cycle progression while E2F4 and E2F5 are constitutively expressed (Crosby & Almasan, 2004).

Several mechanisms are involved in regulation of E2F4 activity: sub-cellular localization, interactions with retinoblastoma proteins (RB), post translational modification such as phosphorylation, and decreased translation mediated by antisense transcripts (Lindeman et al, 1997; Yochum et al, 2007). Unlike E2F1, E2F4 primarily exists in the cytoplasm during cell cycle progression due to lack of a nuclear localization signal (NLS). Upon cell cycle arrest, cells start to enter into G_0, and cytoplasmic E2F4 forms heterodimers with a DP protein, which facilitates the localization of E2F4

complexes into the nucleus in a CRM1 mediated manner (Gaubatz et al, 2001; Moberg et al, 1996). Nuclear localized E2F4 complexes bind to target promoters and regulate expression of diverse classes of genes involved in cell cycle, DNA repair and apoptosis (Balciunaite et al, 2005). Three pocket proteins including pRB, p107/RBL1, and p130/RBL2 are crucial cofactors for the regulation of E2Fs, and their expression level changes during cell cycle (Roman, 2006). Thus, it is believed that it has crucial roles in mediating cell cycle arrest along with RBL2 in G_0 rather than promoting cell cycle progression. The binding of the E2F4-RBL2 complex to E2F4 responsive promoters triggers the recruitment of HDAC complexes or other co-repressors, resulting in the repression of target gene expression (Crosby & Almasan, 2004; Meloni et al, 1999).

Other observations however are not consistent with the view that E2F4 is exclusively a repressor of cell proliferation. Aberrant expression or mutation of E2F4 triggers the malfunction of cell cycle controls and results in malignant tumors. Transfection of E2F4 into non-transformed cells induces the oncogenic activity of E2F4 (Souza et al, 1997). Moreover, overexpression of E2F4 in transgenic mice causes tumors, providing evidence for the oncogenic activity of E2F4 (Wang et al, 2000). Many cancers have mutated E2F4 such as colorectal carcinomas, endometrial cancers, gastric adenocarcinomas, prostatic carcinomas, and ulcerative colitis-associated neoplasms. These facts further emphasize the important role E2F4 plays in tumorigenesis (Schwemmle & Pfeifer, 2000; Souza et al, 1997). The newly discovered function of E2F1-E2F3, switching roles from activators to repressors, suggests that the function of

other members of this family of regulators may be also be more malleable than previously thought (Chong et al, 2009). In order to better understand the physiological roles of E2F4 and reconstruct its regulatory network, it is essential to identify genome-wide E2F4 targets and establish how target promoters respond to it.

Several chromatin immunoprecipitation (ChIP) coupled with chip (ChIP-chip) experiments have been conducted on core-promoter arrays with quiescent and continuously growing cells. These studies have revealed that several hundred E2F4 targets that are involved in diverse functions such as cell cycle regulation, DNA damage repair, apoptosis, mRNA processing, ubiquitination, etc (Balciunaite et al, 2005; Cam et al, 2004). A recent ChIP-chip study of E2F4 using tiled ENCODE arrays identified 187 E2F4 binding sites in 1% of the human genome in lymphoblastoid cells (Xu et al, 2007), suggesting the possibility that E2F4 may have more than 10,000 binding sites across the entire human genome. Moreover, although E2F4 showed a strong binding preference to promoters, some E2F4 binding sites were discovered in non-promoter regions. Without a comprehensive and unbiased genome-wide target analysis of E2F4, it is difficult to evaluate its promoter binding preferences or gain a complete understanding of its functions as a transcription factor.

The recently developed chromatin immunoprecipitation (ChIP) followed by sequencing (ChIP-seq) technique allow us to investigate a genome-wide unbiased search for binding sites of TFs (Barski et al, 2007; Ji et al, 2008; Mikkelsen et al, 2007; Valouev et al, 2008). We used ChIP-seq to catalog E2F4 binding sites across the genome in the

human B-lymphoblastoid cell line, GM06990. We discovered 16,246 putative E2F4 binding sites distributed across promoters to coding and non-coding regions, providing evidence to support diverse roles of E2F4, which were not reported in previous studies. Furthermore, gene expression profiling in response to overexpression of E2F4 in the presence of its cofactors showed that it can function as an activator as well as a repressor of transcription.

The contents of chapter 2 were published in Nucleic Acid Research.
Lee BK, Bhinge AA, Iyer VR (2011) Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. *Nucleic Acids Res*. May 1;39(9):3558-73.

2.2 MATERIALS AND METHODS

Cell culture

The lymphoblastoid (GM06990) cell line was purchased from Coriell and cultivated in RPMI medium containing 15% fetal bovine serum (FBS) as well as 1% antibiotics (penicillin/streptomycin). To perform time course ChIP experiments, cells were harvested every 24 hr for 5 days. Serum starvation was achieved by washing cells cultured for 72 hr three times with RPMI medium without FBS, then adding low-serum RPMI medium containing 0.1% FBS and then cultivating them for 2 days. For serum activation, low serum medium was replaced with RPMI containing 15% FBS. Cells were then cultivated and harvested at 3, 9, and 18 hr.

ChIP sequencing

ChIP assays were performed as described previously (Kim et al, 2008). Briefly, GM06990 cells cultured for 72 hr were cross-linked with 1% formaldehyde and incubated for 7 min at room temperature. Formaldehyde was deactivated by the addition of glycine (125 mM final concentration). Sonicated cell lysate containing an average size of 500 bp DNA fragments was used for immunoprecipitation to enrich E2F4-DNA complexes using an anti-E2F4 antibody (SC-1082X, Santa Cruz Biotech). Immunoprecipitated DNA was sequenced using Illumina sequencing technology (single end sequencing). Data from this study is available at the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/projects/geo/, GSE21488 and GSE21439).

Quantitative-PCR validation

Primer pairs for 42 targets and a negative control region (Appendix A) for normalization were designed using Primer3 (http://frodo.wi.mit.edu/primer3/input.htm). Quantitative-PCR (qPCR) was performed using the SYBR green PCR kit from Applied Biosystems with 1 ng of ChIP and input DNA. Fold enrichment of targets in ChIP DNA relative to input was calculated from an average of three replicate qPCR reactions.

E2F4 overexpression and expression microarrays

Full length E2F4, DP-1, and RBL2 clones were purchased from Open Biosystems and subcloned into the pcDNA 3.1 vector (Invitrogen). Either full length expression constructs or empty vectors as a control were transfected into HeLa cells using Lipofectamine 2000 from Invitrogen. Total RNA was extracted from cells transfected with combinations of the three factors (E2F4, DP-1 and RBL2) or vector transfected cells using Trizol. Microarray experiments were performed using spotted HEEBO oligonucleotide human arrays (Kim & Pollack, 2009), which has 44,308 probes, using the protocol described previously (Gu & Iyer, 2006). Briefly, total RNA was converted into cDNA and labeled with Cy dyes (Cy3 for control and Cy5 for TF overexpression). Dye-coupled cDNA was combined and hybridized onto the oligo arrays for 14 hr. Cy5/Cy3 ratios were calculated from scanned intensity data from each channel. Data were normalized and analyzed by the error model described previously (Hu et al, 2007).

TaqMan assay for miRNA expression

Total RNA was extracted from relevant samples by Trizol. All primer sets for specific miRNAs and PCR reagents for TaqMan miRNA assay were purchased from Applied Biosystems and real time PCR was performed using a 7900HT real time PCR machine from Applied Biosystems. RNU66 was used as an internal control for normalization. miRNA gene expression levels relative to the control was calculated from an average of 4 replicate qPCR reactions.

Luciferase reporter gene assay

Around 700 bp of PCR-amplified insert from each of ten distal binding sites was cloned into the upstream position of a SV40 promoter in a pGL3 plasmid (Promega cat. # E1761) between the KpnI and XhoI restriction sites. All primers used for cloning are listed in Appendix A. For the luciferase reporter gene assay, approximately $3x10^5$ HEK 293 cells were co-transfected with 200 ng of the pGL3 vector or reporter construct containing the Firefly reporter gene as well as 10 ng of pRL-PK vector (Promega cat. # E2241) containing a Renilla reporter gene, which served as an internal control reporter, using 1 µl of lipofectamine 2000 (Invitrogen) and Opti-MEM medium (Invitrogen). After transfection and incubation for 24 hr, cells were washed with PBS once, lysed, and assayed to measure luciferase activity using the Dual Luciferase assay Kit (Promega cat. # E1910) according to manufacturer's instructions. Firefly luciferase activity from the pGL3 construct was then normalized to Renilla luciferase activity. To calculate relative

expression fold change, Firefly activity of the pGL3 vector containing a distal E2F4 binding site was further normalized with that of an empty pGL3 vector. *P*-value was calculated from three independent transfections using a t-test.

Identification of ChIP-seq binding peaks and sites

Illumina sequencing generated 23-32 base pair short reads from the ends of ChIP-enriched DNA fragments. These short reads were mapped back to the genome using the ELAND algorithm. We obtained 6,508,011 uniquely aligning reads from the E2F4 chip library and 8,474,489 uniquely aligning reads from the input library. To identify E2F4 binding sites from high-throughput Illumina sequencing data, we used a Parzen window based algorithm as described previously with minor modifications (Shivaswamy et al, 2008). Each read was assigned a score that was essentially the frequency of observing that read in the sequencing library. The plus strand reads and the minus strand reads were analyzed separately to find peaks on the plus and minus strands respectively. The algorithm begins by assigning the score of each read to its neighboring nucleotides as a function of the read's distance from that nucleotide. The function used to assign scores was a Gaussian kernel with a defined band-width. Local maxima on the plus and minus strands were defined as peaks. High scoring plus peaks that are upstream and within 500 bp of minus strand peaks were considered to be paired and the distance between the paired plus and the minus peak was calculated as the fragment length. A second iteration of peak finding was then carried out, where all aligning reads were extended in the 3'

direction by half of the previously estimated fragment length, to effectively represent the center of the ChIP fragment. The peak-finding algorithm described above was used again on these positions to find local maxima across the genome thereby defining binding sites. The score associated with the nucleotide corresponding to the maxima was assigned to the binding site.

Input correction

In order to correct high ChIP-seq scores arising from repetitive sequences or copy number repeats rather than true ChIP enrichment, we normalized the E2F4 scores by the parallel input sequencing scores. Scores for E2F4 peaks that were within 500 bp from any input peak were divided by the corresponding input peak score. If a given E2F4 peak overlapped with more than one input peak, the higher scoring input peak was used for the correction. E2F4 peaks mapping to within 10,000 bp from any TSS of a gene were not input corrected, since peaks near promoters in sonicated crosslinked chromatin can arise even in input DNA due to transcription factor binding (Auerbach et al, 2009).

False discovery rate

We ran the peak-finding algorithm on a set of randomly simulated read coordinates equal in number to the ChIP-seq data. These simulations were repeated 20 times. At each of a series of different score thresholds, the number of E2F4 peaks found after input correction was compared to those found in the random simulations to give the false discovery rate.

Saturation

We used a capture-recapture analysis to estimate saturation of binding sites in our E2F4 data. Capture-recapture analysis has been used to estimate population sizes of animals in a given area. The reads from the E2F4 chip library were obtained in two sets or "lanes", the first set having 2,305,280 reads and the second set having 4,202,731 reads. Each set was treated as an independent capture. The entire genome was binned into 500 bp bins and reads mapping to each bin were counted for the two sets separately. Each bin was now assigned a p-value value dependent on the number of reads observed within that bin according to a random Poisson model. At different p-value thresholds, we calculated the following:

N1: Number of bins in set 1

N2: Number of bins in set 2

K: Number of bins common to set 1 and set 2

E1: Expected number of bins in set 1 according to a random Poisson model

E2: Expected number of bins in set 2 according to a random Poisson model

FDR1: (E1/N1)*100

FDR2: (E2/N2)*100

E1 and E2 represent the expected number of false positives at each enrichment cut-off.

The estimated number of E2F4 bins at each p-value cut-off was calculated as:

$P = (N1 - E1)(N2 - E2)/ \min[k(1 - E1/N1), k(1 - E2/N2)]$

Whereas the observed number of bins was calculated as:

$O = (N1 – E1) + (N2 – E2) – \min[k(1 – E1/N1), k(1 – E2/N2)]$

The percentage saturation was calculated as:

$S = (O/P)*100$

For each p-value cut-off, we calculated the average false discovery rate as the geometric mean of FDR1 and FDR2. The percentage saturation (S) was now plotted as a function of the average FDR.

Mapping binding sites to gene features

To detect E2F4 target genes, E2F4 sites were mapped to within 2 kb from the TSS of all genes annotated in the RefSeq database. In order to estimate the number of sites mapping to different gene features, it was necessary to assign one site to one and only one gene feature. Since E2F4 has been known to preferentially bind near the TSSs of genes, we used the following hierarchy to assign sites to features: core > upstream > intron > exon > intergenic. Core was defined as 2,000 bp upstream and downstream from the TSS, upstream was defined as greater than 2,000 bp upstream to a maximum of 20,000 bp upstream from the TSS. Binding sites that could not be mapped to within 20,000 bp upstream of any TSS and were not assigned to any intron or exon were termed intergenic. Genes that had E2F4 binding sites within the core were defined as targets.

Mapping binding sites to miRNAs

In order to identify miRNA targets of E2F4, we excluded binding sites that mapped to core promoters, as it was not possible to unequivocally assign such sites to the

annotated gene or the miRNA using binding data alone. We included miR-22 as a special case for further characterization because we have identified a role for miR-22 in the cell cycle (unpublished data). Sites mapping to intergenic regions, introns and exons were mapped to within 10,000 bp of the annotated starts of mature miRNAs. The data for mature miRNA start/stop coordinates was downloaded from miRBase (www.mirbase.org).

Generating TSS/TTS profiles

A region of 20 kb around the TSS (10 kb upstream and 10 kb downstream), was binned in 50 bp size bins and E2F4 sites were mapped to each bin. Each bin was assigned the score of the peak that mapped to it. Corresponding bin scores were averaged across all genes to generate an average peak score profile across the TSS. This average profile was smoothened by a moving window of 3 bins. The same procedure was used to generate TTS profiles.

Motif analysis

Since attempting to run motif discovery algorithms on all binding sites would have been computationally expensive, we divided the binding sites into strong (score $\geq$ 24.93), moderate (scores 8.01 to 9.01) and weak (scores 5.6 to 5.75) categories and considered the top 500 sites from each category for motif discovery. A 200 bp region centered on each site was extracted from the human genome assembly hg18. Motif discovery was performed using the software DRIM (Eden et al, 2007) on each category

separately at a p-value threshold of 1×10^{-5}. A random background was generated by sampling 200,000 sequences of 200 bp from the genome. 55.9% of E2F4 sites occurred within 2 kb upstream and downstream of TSSs of genes and this ratio was maintained in the random sample. In addition, we calculated the enrichment of each motif with respect to the random background as a function of the peak score. To analyze the relationship between each motif and different gene features like core, upstream, intron, exon and intergenic (as defined above), we divided sites into different features such that each site was assigned to one and only one feature (as described above) and then extracted sequences associated with these feature-specific sites (200 bp centered on the site). Then, for each feature, we counted the number of sites that had a given motif and divided this number by the total number of sites that had that specific motif. This was repeated for each of the 5 analyzed features for a given motif and the data was displayed as a heat-map.

Motif co-enrichment

Conserved TF binding site data for the human genome assembly hg18 was obtained from http://genome.ucsc.edu. This data contains TF bind sites (TFBS) for 398 TFs from the TRANSFAC database that are conserved between human, mouse and rat (http://genome.ucsc.edu/cgi/bin/hgTrackUi?hgsid=118849564&c=chr13&g=tfbsConsSites). A TFBS was considered associated with an E2F4 site if it was found within 250 bp of that site. The frequencies of TFBS associated with binding sites were calculated for E2F4

peaks as well as for the randomly generated peaks. The analysis was performed on the strong, moderate, and weak categories separately and p-values were calculated according to a binomial model. We excluded TFBS that were not enriched at a p-value of $< 1 \times 10^{-6}$ and were associated with less than 4% of the E2F4 sites under consideration.

Motif co-occurrence

We counted the number of different motifs that were associated with each binding site peak (that is, within 100 bp on either side of the peak), and compared the distribution of these counts between randomly generated peaks and E2F4 binding site peaks.

2.3 RESULTS

Optimal cell culture conditions for E2F4 ChIP in lymphoblastoid cells.

Since E2F4 is reported to bind its target promoters in quiescent cells, we first investigated E2F4 ChIP efficiency in lymphoblastoid cells that were grown for 24, 48, 72 and 96 hours after replating by hybridizing the ChIP-enriched DNA to a core-promoter array that covered -750 bp to +200 bp from the transcription start site (TSS) of ~9000 genes. We also examined E2F4 occupancy onto its target promoters in serum-starved as well as serum-fed cells on core promoter arrays because it is known that upon serum starvation, E2F4 accumulates in the nucleus and is recruited to its target promoters, resulting in cell cycle arrest (Deschenes et al, 2004). Strong occupancy signals for E2F4 were observed at 72 hr and were maintained at 96 hr. We found that the E2F4 binding profile was not significantly affected by serum-starvation or stimulation in lymphoblastoid cells (Fig. 2-1). Based on the above results, we generated ChIP-seq data for E2F4 in lymphoblastoid cells that were maintained in culture for 72 hrs.

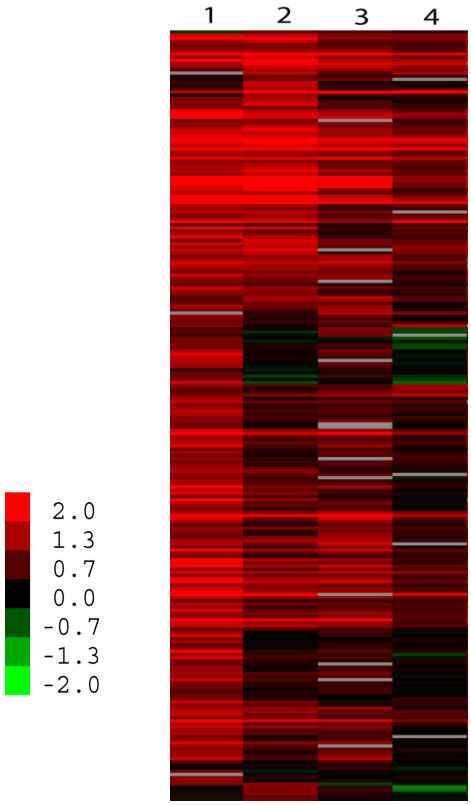

Figure 2-1. Time-course ChIP-chip experiments for E2F4 on core promoter arrays.
(1) 72 hr culture, (2) 96 hr culture, (3) serum starvation for 72 hr, (4) serum activation for 3 hr. The ratio of ChIP to input determined by microarray hybridization is plotted using the color scale shown.

Identification and verification of E2F4 binding sites from ChIP-seq data

ChIP-seq of E2F4 generated around 6.5 million uniquely mapped reads to the reference human genome. In order to identify E2F4 binding sites from this data, we developed a peak detection program using a Parzen windows density estimation algorithm we have previously used to map nucleosome positions (Shivaswamy et al, 2008). We also sequenced an input DNA corresponding to ChIP DNA from the same

cells as a control to ensure that enriched sites were not an artifact of the processing and sequencing. Our algorithm identifies discrete regions of approximately 150 bp around each ChIP-seq peak which we refer to as sites. Known E2F4 targets were easily detected in our ChIP-seq data (Fig. 2-2A). We observed that some strong peak in the ChIP dataset corresponded to equally strong peak in the input sequencing data (Fig. 2-2B). Such peaks may arise due to the presence of repeat regions in the genome and are thus false positives.

To minimize such false positive targets, we first normalized the ChIP peaks by dividing the ChIP peak scores with their corresponding input peak scores. Next, we calculated a false discovery rate (FDR) based on random simulations to decide an appropriate significance threshold (Fig. 2-2C). At 1% FDR, which corresponded to an input-corrected peak score of 4.4, we identified 16,246 putative E2F4 binding sites across the entire genome. To verify the quality of the putative E2F4 binding sites, we examined overlaps between ChIP-seq targets with those from our core promoter arrays. About 84% of core promoter targets were also overlapped with those of the ChIP-seq data. We further verified our ChIP-seq data by performing quantitative-PCR (qPCR) for 42 randomly selected targets, including 30 targets with a score between 4.4 and 8, as well as 12 targets with a score less than 4.4, which was below our 1% FDR threshold. Overall, binding sites with stronger ChIP-seq scores showed higher fold enrichment by qPCR. Specifically, more than 90% of the targets above the 1% FDR threshold showed an enrichment of at least 1.5 fold by qPCR. On the other hand, only 41% of sites below the 1% FDR threshold showed enrichment by qPCR (Fig. 2-2D).

We also estimated the extent to which we identified all E2F4 sites in the genome, using a tagging-recapture saturation analysis (see Methods). At the FDR threshold of 1% and our given sequencing depth, we estimated that we identified more than 80% of all E2F4 binding sites in the human genome in lymphoblastoid cells (Fig. 2-2E).

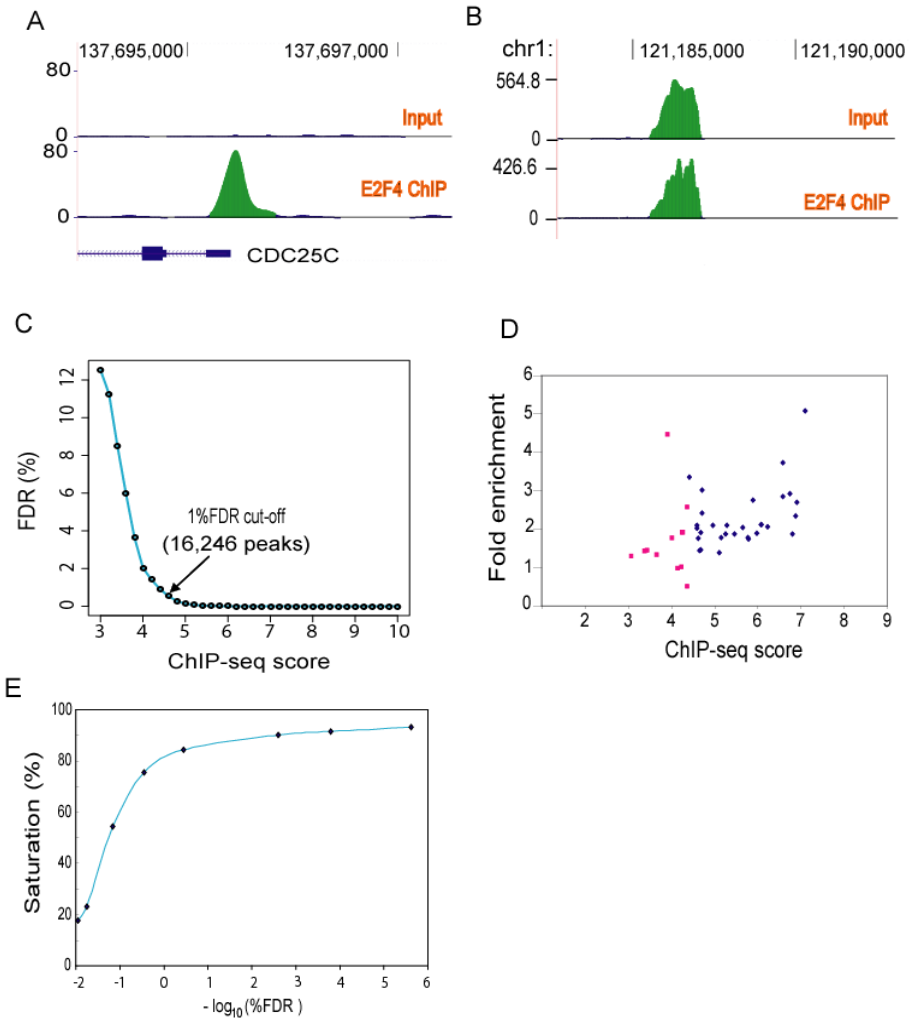

Figure 2-2. E2F4 ChIP-seq reveals genome-wide E2F4 binding sites.
(A) An example of a known E2F4 binding site that was identified in our ChIP-seq data. Chromosome coordinates are indicated on top. The plot in the middle shows the density of ChIP-seq reads, with the peak score indicated on the Y axis. The bottom track shows the CDC25C gene with coding regions, exons and introns indicated by thick or thin boxes and line respectively. The direction of transcription is indicated by the arrows from right to left. (B) An example of strong peaks discovered in both input and ChIP likely due to copy number differences between the cell genome and the reference sequence. Such sites were removed by input correction (see Methods). (C) FDR calculation based on random simulations. The 1% FDR threshold was used for further analysis. (D) Quantitative PCR verification of 42 randomly selected targets identified by ChIP-seq. Blue diamonds represent targets which passed the 1% FDR ChIP-seq threshold, and red squares represent targets below this threshold. (E) Capture-recapture analysis to estimate saturation for E2F4 targets (see Methods). X-axis represents - $\log_{10}$ FDR and the Y axis shows the saturation as a percentage of expected sites that were discovered at each FDR.

Distribution of E2F4 binding sites in relation to gene annotations

We first investigated the relationship between E2F4 binding sites with gene density. E2F4 binding sites are positively correlated with gene density across the genome (r^2 = 0.75) (Fig. 2-3A). Next, we examined the distribution of E2F4 binding sites in 5 different genomic regions including promoter, exon, intron, intergenic and upstream regions by mapping E2F4 sites relative to RefSeq annotated genes. Approximately 56% of the sites occurred within promoters (Fig. 2-3B). In addition, the binding profile of E2F4 around TSSs also provided evidence that E2F4 had a preference for binding promoters, especially near the TSS, whereas no significant binding preference was observed near the transcription termination sites (TTSs) (Fig. 2-3C). This result is consistent with previous studies using selective arrays, showing that the binding sites of several TFs, including E2F1, were mainly distributed near the TSSs (Ren et al, 2002; Tabach et al, 2007; Xu et al, 2007).

Since E2F4 showed preferential binding at promoters, we examined whether the binding strength as measured by the peak score of E2F4 sites at promoters was stronger than sites at other genomic regions. E2F4 promoter sites showed significantly higher peak scores compared to those from other genomic regions (Fig. 2-3D). Taken together, these results show that not only does E2F4 bind preferentially to promoters, but it also binds with higher occupancy to promoters as compared to other genomic regions.

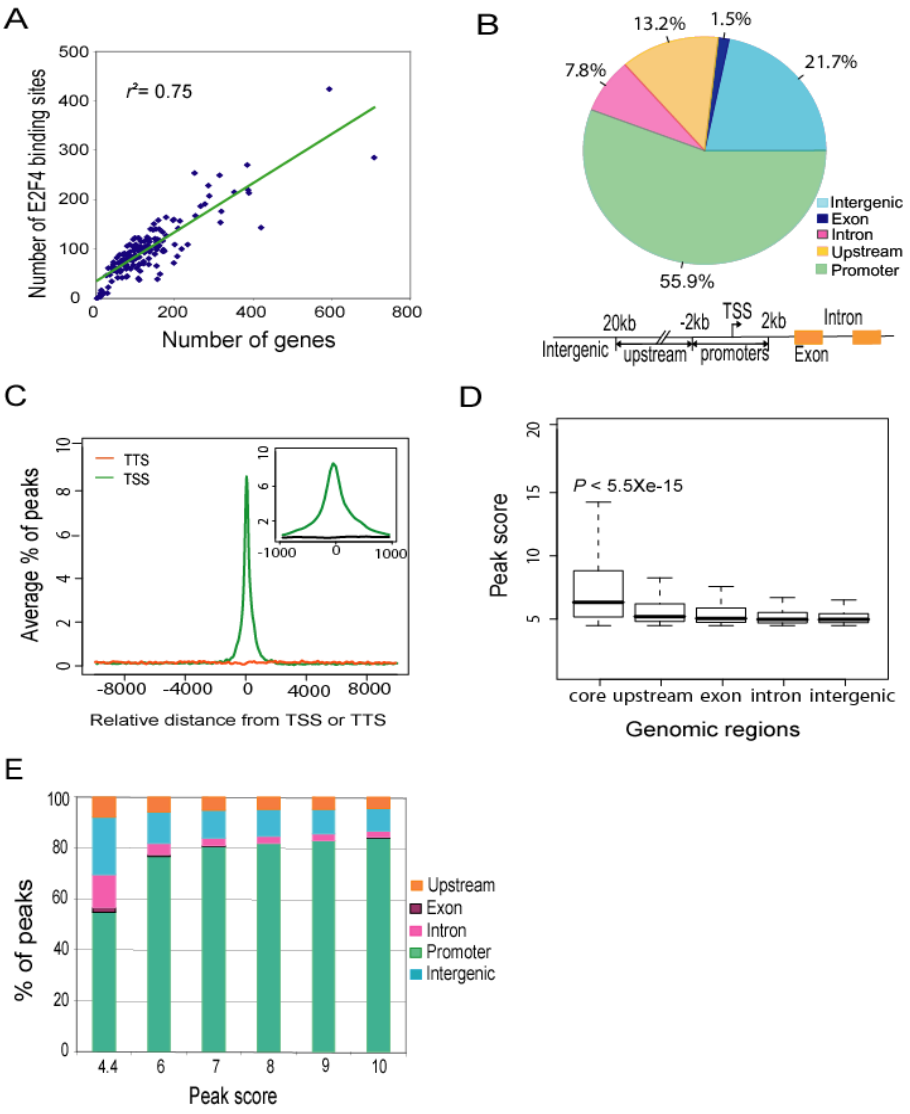

Figure 2-3. The genome-wide distribution pattern of E2F4 binding sites.
(A) The correlation between E2F4 binding sites and gene density. Each point on the plot represents a 20 Mb bin. (B) A pie chart representation of the distribution of E2F4 binding sites in 5 different genomic regions. The definition of each genomic region is described below. Core promoters are within ± 2 kb from the TSS, upstream is from 2 kb to 20 kb upstream from the TSS, and intergenic is a region not included as a promoter, upstream region, intron or exon. (C) Distribution of E2F4 binding sites within ± 10kb. Inset shows a close up of a 1 kb region centered on the TSS. (D) A box-plot shows the ChIP-seq peak score distribution across 5 different genomic regions. Peak scores in core promoters were significantly higher compared to those from other genomic regions ($P < 5.5 \times 10^{-15}$, Wilcoxon test with Bonferroni correction). (E) Distribution of E2F4 binding sites depending on peak scores. Even though the number of intergenic sites decreased with increasing score, a substantial proportion of intergenic sites (10%) still remained at a score of 10.

E2F4 and bidirectional promoters

It has been reported that approximately 11% of all genes have bidirectional promoters in mammalian genomes (Adachi & Lieber, 2002; Trinklein et al, 2004). Recent computation analysis has also reported that consensus E2F4 motifs are significantly overrepresented in bidirectional promoters (Lin et al, 2007). In order to examine whether E2F4 binding sites showed a bias toward binding to bidirectional promoters, we first defined bidirectional promoters as the region of DNA between the TSSs of two genes that were divergently transcribed from opposite strands and separated by less than 2 kb. Based on this criterion, we identified 918 bidirectional promoters corresponding to 1836 genes (9.8%) among the 18,693 genes annotated in RefSeq. We mapped E2F4 sites to these bidirectional promoters and found 572 (31%) genes to be bound by E2F4 (Appendix B), which was a significant enrichment over background (hypergeometric $P < 1.1\text{x}10^{-10}$), indicating that many divergently transcribed genes in the genome might be co-regulated by E2F4. Divergently transcribed E2F4 target genes were highly overrepresented in the categories of RNA processing, DNA repair, protein folding, and cell cycle.

Distal E2F4 sites could be enhancers or other regulatory elements

In addition to strong promoter occupancy and bidirectional promoter enrichment of E2F4 binding sites, a proportion of E2F4 sites were also detected in introns (7.8%), upstream regions (13.2%), and intergenic regions (21.7%) (Fig. 2-3B). Previous

approaches have not identified this latter class of E2F4 sites that are not at the core promoter. To exclude the possibility that the limited number of TSS annotated in Refseq (~18,000) was resulting in an overestimate of the number of intergenic binding sites, we mapped all E2F4 sites to an expanded data set of approximately 60,000 TSSs derived by combining RefFlat annotated genes with additional gene annotations obtained from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables) and further filtered to remove redundant TSS coordinates. Even with the much larger number of TSSs used in this analysis, the number of intergenic E2F4 sites showed only a modest decrease from 22.5% to 17.5%, indicating that a significant proportion of E2F4 sites are truly intergenic. We then analyzed the distribution of E2F4 sites at different peak score cut-offs to investigate whether the percentage of intergenic sites was dependent on the score. Although the proportion of intergenic sites decreased with an increase in the score threshold, it remained fairly constant above a score cut-off of 10, where approximately 10% of E2F4 sites were deemed intergenic (Fig. 2-3E). Overall, these results indicate the existence of a significant number of strong E2F4 sites that were found greater than 20 kb away from any annotated TSS.

It is well established that TFs are able to regulate the expression of target genes by binding to promoters or long-range regulatory elements such as enhancers and insulators. To investigate the possibility that some of the distal (9.6% of upstream and 17.5 % of intergenic) E2F4 binding sites represent enhancers, we examined these distal E2F4 binding sites for the presence of characteristic enhancer signature marks based on

published histone modification data (Barski et al, 2007; Heintzman et al, 2007; Wang et al, 2008). Genome-wide histone signature analyses have revealed that the three forms of H3K4 methylation (H3K4me1, H3K4me2, and H3K4me3), H3K9me1, H3K18ac, and the variant H2A.Z are highly enriched in enhancer sites. We found that 36% of the intergenic E2F4 binding sites and 68% of the upstream sites had at least one histone enhancer marker, with most of these E2F4 sites showing multiple enhancer makers (Fig. 2-4A). To verify the possibility that distal E2F4 binding sites could function as enhancers, we first tested whether they showed binding by p300, a known marker of enhancers in mammalian cells. We tested p300 binding at 10 randomly selected distal E2F4 enhancer candidate sites using ChIP followed by real-time PCR. We found that 9 out of 10 sites showed significant binding by p300 (Fig 2-4B), which supported the possibility that some of the distal E2F4 binding sites we identified by ChIP-seq could function as enhancers.

To functionally test this possibility further, we cloned these 10 candidates into pGL3 promoter-containing enhancer reporter vectors and performed luciferase reporter gene assays. Five of the 10 sites showed significant increase in expression of their reporter genes ($P < 0.005$), which confirmed that a subset of distal E2F4 binding sites could as enhancers (Fig. 2-4C). Based on the reporter assays, we can roughly estimate approximately 1,048 out of 4147 distal E2F4 binding sites may function as enhancers. Distal E2F4 sites that did not show any enhancer marks may potentially regulate non-coding genes such as miRNAs whose promoters are not well defined, or these distal sites could be transcriptionally neutral.

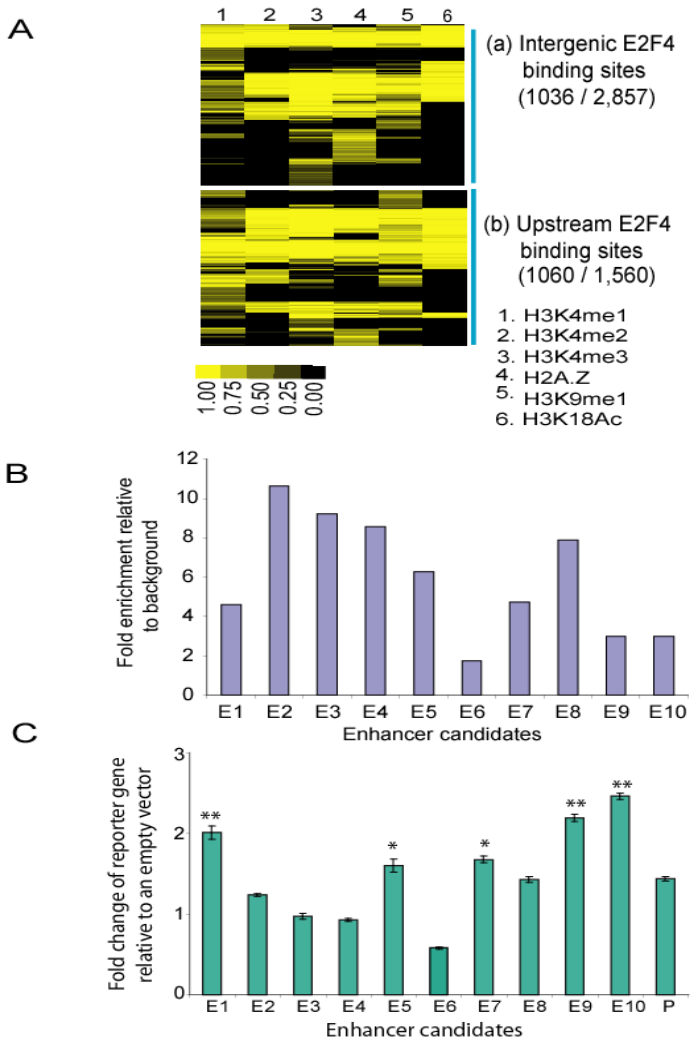

Figure 2-4. Some E2F4-bound distal sites function as enhancers.
(A) Relationship of histone enhancer marks with the 2,857 intergenic E2F4 sites (a) or 1560 upstream E2F4 sites (b). Data for histone modifications indicative of enhancers was obtained from Barski et al. 2007 and Wang et al. 2008, assigned to E2F4 binding sites identified here and hierarchically clustered for display. The relative strength of the histone modification signal is indicated in the heat-map according to the indicated color table. 68% of upstream and 36% of intergenic E2F4 sites contains at least one enhancer mark. (B) Fold enrichment of p300 binding for 10 randomly selected enhancer candidates from E2F4 ChIP-seq data. (C) Luciferase reporter gene assays for randomly chosen 10 distal E2F4 binding sites. The Y-axis represents the expression fold change of a luciferase reporter gene normalized to an empty-vector control. P-values were calculated using t-test from three independent transfections. E1 through E10 represent 10 enhancer candidates randomly selected from among distal E2F4 binding sites. P represents a positive control enhancer selected from (Heintzman et al, 2009). One and two asterisks mean $P < 0.005$ and $P < 0.001$, respectively.

Putative E2F4 target genes are involved in a broad range of biological processes.

In order to investigate the functions of E2F4 suggested by its genome-wide binding profile, we first considered all E2F4 sites that occurred within ± 2 kb from the TSS of all genes annotated in the RefSeq database. We found 7,346 genes that had E2F4 binding sites in their promoters, which cover almost 30% of all annotated human genes.

Next, we analyzed functional categories among these putative E2F4 target genes using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis et al, 2003). In agreement with previously reported results, E2F4 target genes were highly enriched for cell cycle, DNA repair, RNA processing, stress response, apoptosis, and ubiquitination (Table 2-1). We also found significant enrichment among E2F4 targets for additional functions that have not previously been associated with E2F4, such as protein transport and targeting, protein folding, and I-kappaB kinase/NF-kappaB cascade. KEGG pathway analysis also showed strong enrichment of E2F4 in the categories of cell cycle, ubiquitin-mediated proteolysis, p53 signaling pathway and chronic myeloid leukemia.

We also found that E2F4 binds to the promoters of 780 TFs out of the ~2000 known TFs in the human genome (Appendix C), which suggests that E2F4 regulates broad classes of genes indirectly. Taken together, these results suggest that E2F4 could regulate more diverse biological processes than previously suspected.

Table 2-1. Functional categories of E2F4 target genes.

Count represents the number of genes in the biological function category. Percent (%) shows the proportion of E2F4 targets among the count. FDR is the false discovery rate.

Biological functions	Count	%	Fold Enrichment	*P*-value	FDR
biopolymer metabolic process	2259	34.81%	1.36	6.32E-103	1.21E-99
cell cycle	490	7.55%	1.77	8.05E-52	1.54E-48
RNA processing	277	4.27%	1.96	1.55E-39	2.97E-36
organelle organization and biogenesis	579	8.92%	1.57	3.36E-39	6.42E-36
response to DNA damage stimulus	208	3.21%	2.07	7.92E-35	1.51E-31
mRNA processing	172	2.65%	2.14	3.14E-31	6.01E-28
RNA splicing	156	2.40%	2.22	3.23E-31	6.18E-28
protein modification process	772	11.90%	1.38	1.98E-29	3.78E-26
DNA repair	171	2.64%	2.07	1.48E-28	2.83E-25
ubiquitin cycle	278	4.28%	1.74	7.46E-28	1.43E-24
response to endogenous stimulus	230	3.54%	1.84	1.38E-27	2.64E-24
macromolecule localization	401	6.18%	1.54	2.09E-25	4.00E-22
protein transport	345	5.32%	1.6	2.64E-25	5.05E-22
post-translational protein modification	653	10.06%	1.39	4.68E-25	8.96E-22
transcription	1061	16.35%	1.27	1.67E-24	3.19E-21
protein localization	373	5.75%	1.53	1.34E-22	2.56E-19
DNA replication	145	2.23%	1.87	1.42E-18	2.73E-15
apoptosis	356	5.49%	1.47	3.15E-18	6.02E-15
chromatin modification	118	1.82%	1.9	1.11E-15	2.12E-12
DNA packaging	166	2.56%	1.67	1.48E-14	2.85E-11
chromosome segregation	50	0.77%	2.56	2.65E-14	5.05E-11
ribonucleoprotein complex biogenesis and assembly	116	1.79%	1.85	2.84E-14	5.44E-11
protein targeting	122	1.88%	1.8	7.12E-14	1.36E-10
cell development	489	7.54%	1.27	1.16E-10	2.21E-07
RNA localization	58	0.89%	2.1	1.36E-10	2.61E-07
sister chromatid segregation	28	0.43%	2.82	6.29E-10	1.20E-06
protein ubiquitination	51	0.79%	2.11	1.76E-09	3.37E-06
ribosome biogenesis and assembly	55	0.85%	1.95	2.16E-08	4.14E-05
protein kinase cascade	174	2.68%	1.43	2.96E-08	5.67E-05
response to stress	416	6.41%	1.24	6.16E-08	1.18E-04
spindle organization and biogenesis	19	0.29%	3.06	6.67E-08	1.28E-04
phosphate metabolic process	399	6.15%	1.25	1.07E-07	2.04E-04
phosphorus metabolic process	399	6.15%	1.25	1.07E-07	2.04E-04
protein folding	127	1.96%	1.49	2.01E-07	3.85E-04
DNA damage response, signal transduction	33	0.51%	2.22	3.98E-07	7.62E-04
protein-RNA complex assembly	62	0.96%	1.73	1.07E-06	0.002
regulation of gene expression, epigenetic	34	0.52%	2.11	1.43E-06	0.002
chromatin assembly or disassembly	73	1.12%	1.63	2.06E-06	0.003
lipid biosynthetic process	124	1.91%	1.44	2.09E-06	0.004
microtubule organization and biogenesis	17	0.26%	2.89	2.53E-06	0.004
centrosome organization and biogenesis	17	0.26%	2.89	2.53E-06	0.004
I-kappaB kinase/NF-kappaB cascade	69	1.06%	1.63	3.98E-06	0.007

E2F4 potentially regulates other E2F family members and its cofactors.

Previous ChIP-chip studies have revealed that members of the E2F family transcriptionally regulate each other (Balciunaite et al, 2005; Ren et al, 2002). For instance, E2F4 occupies the promoters of activator E2Fs (E2F1, E2F2, and E2F3) and represses their expression to cause cell-cycle arrest. We found that E2F4 occupied the promoters of all E2F family genes including E2F7 and E2F8, which are RB independent repressors (Rowland & Bernards, 2006). The notable exceptions were E2F4 itself, and E2F6 (Table 2-2), which interestingly, functions redundantly with E2F4 as a repressor. E2F4 also occupied the promoters of the three RB family proteins (pRB, p107/RBL1, and p130/RBL2), and two binding partners (DP1 and DP2). In particular, E2F4 showed strong binding at the promoters of E2F1-E2F3, which are genes involved in cell cycle progression. To identify targets that were common to E2F1 and E2F4, we compared our E2F4 targets with previously published E2F1 targets obtained from ChIP-chip data in the same cell line (Xu et al, 2007). Among the top 2,000 known E2F1 targets, 1,416 targets (~70%) overlapped with our E2F4 targets identified by ChIP-seq. Functional analysis revealed that cell cycle and DNA repair functions were highly enriched among these genes that were occupied by both E2F1 and E2F4.

The retinoblastoma (RB) protein family has important roles in the regulation of E2F activity. Several studies have showed that E2F4 can form a complex with one of three RB proteins, and that the abundance of the E2F4-RB complex varies depending on the cell cycle state (Cam et al, 2004; Moberg et al, 1996). We compared the overlap of

our E2F4 targets with previously known RBL1 and RBL2 targets identified by ChIP-chip using a core-promoter array of 14,000 genes (Balciunaite et al, 2005). We found that approximately 87% of previously identified RBL1 and RBL2 targets were also bound by E2F4 in our ChIP-seq data. Additionally, we found that 75% of genes previously reported as being bound RBL1 alone were in fact occupied by E2F4 in our dataset, and the majority of the remaining 25% of "RBL1-alone" targets contained E2F4 sites in their promoters that were just below our 1% FDR threshold. This suggests that almost all RBL1/p107 and RBL2/p130 targets are in fact also occupied by E2F4 (Table 2-3).

Table 2-2. E2F4 targets in E2Fs family and their cofactors.

E2Fs and their cofactors	# of reads	score	alias
RB1	19	8.95	pRB
RBL1	74	41.86	p107
RBL2	20	11.57	p130
DP-1	41	17.55	
DP-2	23	7.58	
E2F1	52	23.47	
E2F2	52	33.85	
E2F3	71	36.8	
E2F4	**0**	**0**	
E2F5	8	4.99	
E2F6	**5**	**3.71**	
E2F7	80	35.06	
E2F8	53	26.58	

Table 2-3. Overlap of E2F4 targets with its cofactors.

Gene	# of E2F4 target from Balciunaite et al. 2005.	# of E2F4 target from Chip-seq	% overlap
E2F4(G1)	299	267	89.30
p130(G1)	227	202	88.99
p107(G1)	244	216	88.52
E2F4(G0)	266	238	89.47
p130(G0)	364	318	87.36
E2F4_all	357	314	87.96
p130_all	383	333	86.91

Motif analysis of E2F4 binding sites

E2F family proteins bind to DNA as a heterodimeric E2F-DP complex to the motif TTTc/gGCGCc/g (Zheng et al, 1999). We examined the presence of the consensus E2F motif (TTTSSCGC) over all E2F4 binding sites identified by ChIP-seq. Interestingly, we found that only 5% of E2F4 sites contained the consensus motif, suggesting that E2F4 might be recruited to its sites either through a novel motif or via interaction with other proteins. In order to discover alternative E2F4 motifs, we performed a *de novo* motif search using the DRIM algorithm (Discovering Rank Imbalanced Motifs) (Eden et al, 2007). We first classified E2F4 sites into three different groups, namely strong, moderate, and weak, based on binding strength. Next, we extracted the top 500 sites from each group and then executed DRIM on each set separately. We found a total of 5 different motifs that were significantly enriched over background ($P < 1x10^{-5}$) (Fig. 2-5). Of these, Motif 2 and Motif 3 were similar to motifs recently identified using microarray based in vitro binding experiments for mouse E2F2 and E2F3 (Badis et al, 2009).

To investigate motif occurrence and enrichment, and their dependence on binding strength, we mapped all 6 motifs (1 canonical and 5 newly discovered) back to all E2F4 sites. Fig. 2-5 represents the enrichment of each motif over background as a function of ChIP-seq peak score. The canonical motif (Motif 1) and motif 2 showed the strongest enrichment over background indicating that these two motifs correspond to high occupancy E2F4 binding sites. Motifs 3 and 4 showed moderate enrichment, and motifs 5

and 6 showed weak enrichment over background. All motifs except motif 4 showed a gradual increase in the enrichment as well as in the percentage prevalence amongst binding sites as a function of peak score. Motif 4 hit the highest enrichment around score 10 and showed an apparent decrease of enrichment over background at higher ChIP-seq peak scores. This was mainly because only a small number of high scoring E2F4 sites had this motif. Additionally, we found that for most binding sites, at least one of the 6 different E2F4 motifs was found less than 20 bp from our estimated peak position (Fig. 2-6A).

To investigate whether different motifs are used to recruit E2F4 to different genomic regions, we examined the percentage occurrence of all 6 motifs at five different genomic regions (promoters, upstream, intergenic, introns, and exons) to investigate any regional binding bias of each motif. Most motifs, with the exception of motif 6, were highly overrepresented in promoters, consistent with the occurrence and scores of peaks in these 5 genomic regions (see Fig. 2-3B and Fig. 2-3D). Motif 6 was distinct in that it showed comparable enrichment in promoters and intergenic regions as well as in introns, but not in upstream and exons (Fig. 2-6B), suggesting that motif 6 may have distinct regulatory roles in intergenic and intronic regions. We also examined the number of motifs within each binding site. On average, each E2F4 site contained 2 motifs, while sites with higher scores contained more than 2 motifs (Fig 2-6C), suggesting the possibility of either multiple E2F4 DNA interactions per regulatory region, or usage of distinct motifs under different physiological conditions.

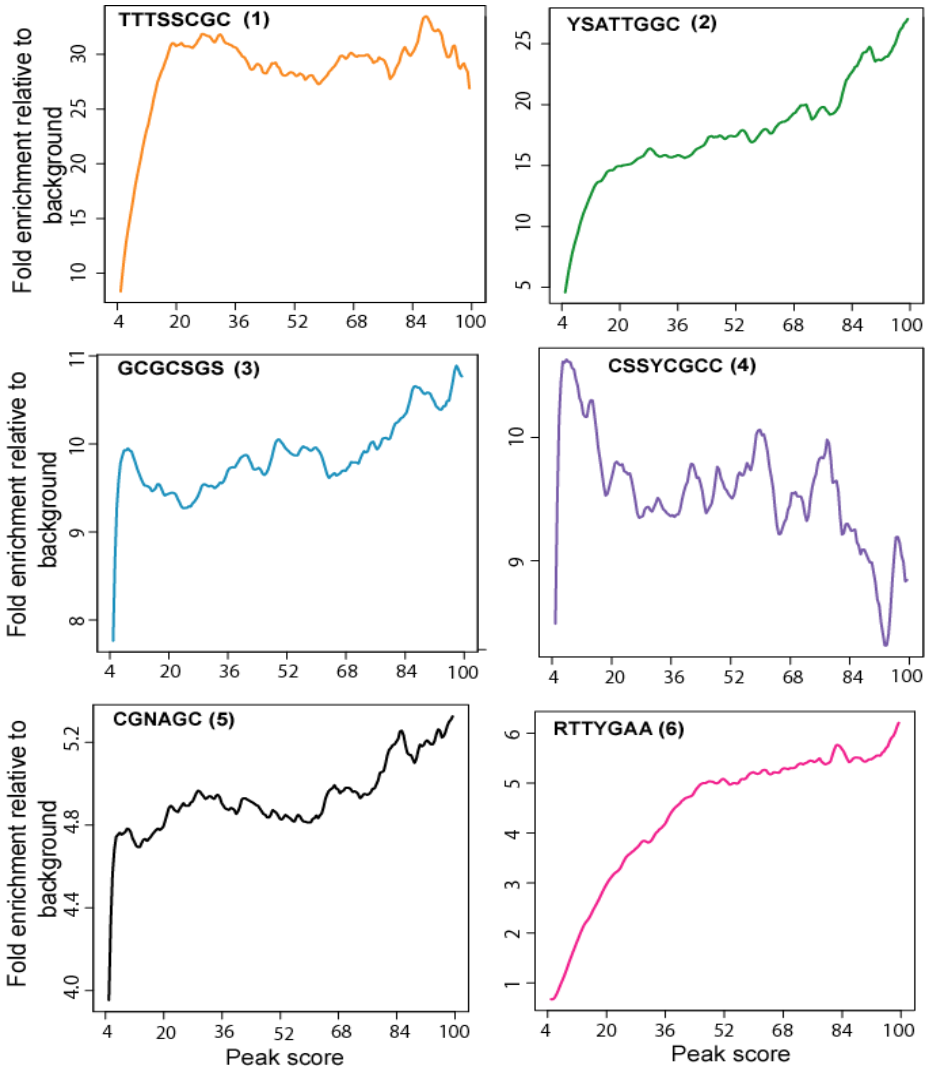

Figure 2-5. Enrichment of indicated motifs over background is plotted on the Y axis, as a function of ChIP-seq peak score plotted on the X axis.

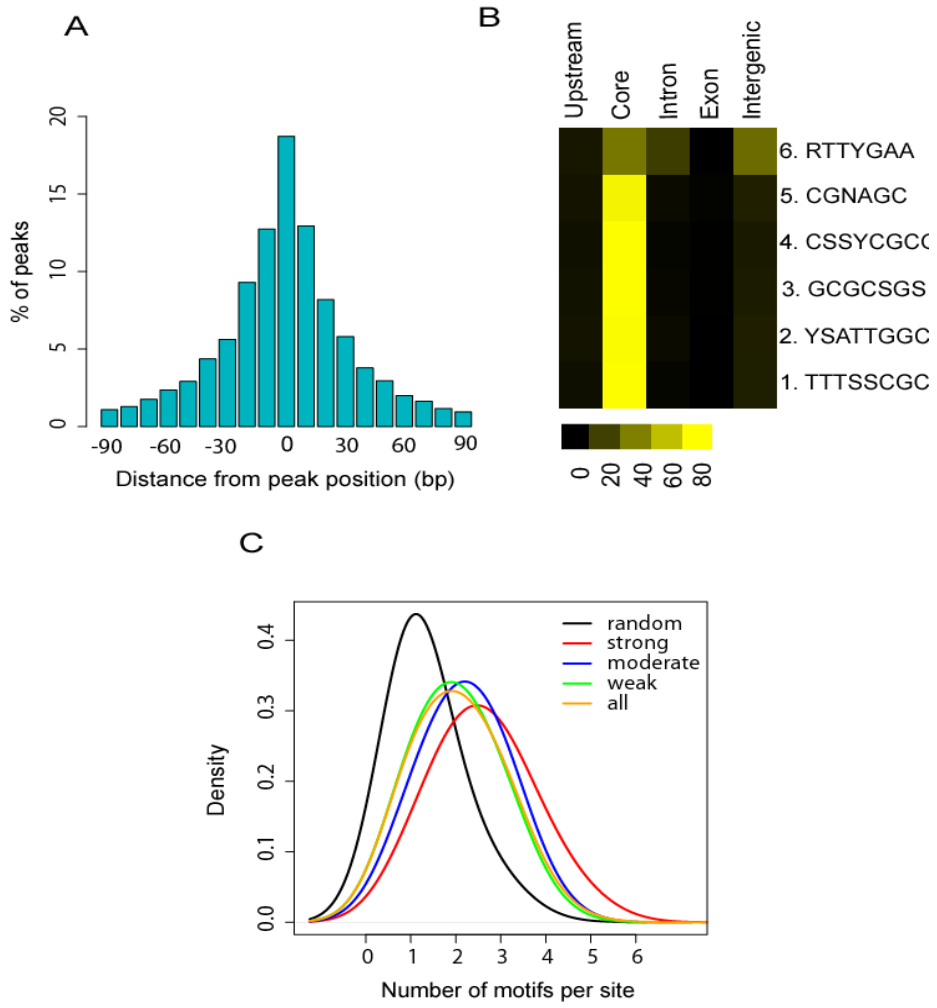

Figure 2-6. E2F4 motif analysis.
(A) Distribution of motifs around E2F4 binding sites identified by ChIP-seq. E2F4 motifs were mapped to E2F4 binding sites and the distance of the identified motif from the maxima of the binding site was plotted as a histogram. The Y-axis shows the percentage of peaks that had an E2F4 motif within the specified distance shown on the X-axis. The figure indicates that the majority of E2F4 peaks had an E2F4 motif within 20 bp of the indicated nucleotide that was designated as the binding site. (B) Frequency of motif occurrence in 5 different genomic regions. The heat-map shows the percentage distribution of E2F4 binding sites found in each genomic region for each of the 6 different E2F4 motifs used in this study. E2F4 motifs 1-5 were found predominantly in sites that mapped to the core promoter except motif 6. Motif 6 was found at almost equal frequency in sites that mapped to the core and intergenic regions. Color bar indicates % of a motif in a given genomic region. For a given motif, the sum of the percentages across all 5 different genomic regions is 100%. (C) Number of motifs discovered within E2F4 sites segregated by their ChIP-seq score. The density plot shows the relative frequency of sites on the Y axis containing each indicated number of motifs on the X axis. Sites with stronger ChIP-seq scores had more motifs and overall, E2F4 sites had approximately 2 motifs per site on average.

In order to investigate whether specific binding motifs were associated with specific functions, we grouped genes based on the presence of specific motifs in their promoters and performed KEGG pathway analysis for each group (Table 2-4). All E2F4 motifs were used to regulate the cell cycle pathway and most motifs were used in several different pathways. However, we found that some pathways were significantly enriched with only one motif. For instance, motif 2 was overrepresented in the biosynthesis of steroids pathway while motif 3 was enriched in the N-glycan biosynthesis pathway, and motif 4 in the chronic myeloid leukemia pathway. Additionally, we found that motif 6 was associated exclusively with cell cycle genes. These results suggest that E2F4 may use distinct motifs to perform specific physiological functions.

Table 2-4. Motif usage of E2F4 within different biological pathways.

Each number indicates one of six E2F4 motifs, assigned to a KEGG pathway category.

KEGG pathway terms	Motifs
Cell cycle	1, 2, 3, 4, 5, 6
Ubiquitin mediated proteolysis	3, 4, 5
Pyrimidine metabolism	1, 3, 5
DNA polymerase	1, 3, 5
p53 signaling pathway	3, 5
Chronic myeloid leukemia	4
N-Glycan biosynthesis	3
Biosynthesis of steroids	2

$P < 1.0$xE-5

We found several other TF motifs to be significantly co-enriched within 500 bp of E2F4 binding sites (Table 2-5). Among them, 10 TFs (EGR1-3, ELK1, PAX5, RFX1, SP1, STAT1, RFX1, and YY1) were also targets of E2F4 in our ChIP-seq data. Many cell cycle progression-related TFs such as AP1, MAZ, ELK1 were also highly enriched in the neighborhood of E2F4 binding sites and such TFs may regulate genes along with E2F4 in a combinatorial or competitive manner.

Table 2-5. Significantly co-enriched transcription factors with E2F4

AHRARNT	AP2	AP4	ARNT	ATF	ATF6	CREBP1
E2F	EGR1	EGR2	EGR3	ELK1	ER	MAZR
MAX	NF1	NF-Y	NRSF1	P53	PAX2	PAX5
RFX1	SP1	SREBP1	STAT1	STAT3	TAXCREB	USF
XBP1	YY1					

($P < 1 \times 10^{-14}$)

Overexpression of E2F4 and its cofactors reveal that E2F4 functions as an activator and a repressor.

It has been reported that siRNA-mediated E2F4 knock-down leads to drug- or irradiation-induced apoptosis (Crosby & Almasan, 2004) while E2F4 knock-down in T98G cells does not affect gene expression due to its functional redundancy with E2F5 and E2F6 (Balciunaite et al, 2005). In order to identify genes whose expression levels are affected by E2F4 and address whether E2F4 functions as an activator or a repressor, we

perturbed E2F4 expression levels by transient overexpression and analyzed gene expression using microarrays. We initially tried overexpressing E2F4 in lymphoblastoid cells; however, the transfection efficiency was too low to discriminate overexpression effects given the background of untransfected cells. We therefore used HeLa cells for gene expression profiling. Before performing overexpression experiments, we compared our E2F4 binding targets from lymphoblastoid cells with targets from HeLa cells obtained from previously published data (Xu et al, 2007) to confirm that E2F4 binding profiles were comparable between the two cell lines. About 81% of the E2F4 targets from HeLa cells were also found in lymphoblastoid cells (Figure 2-7), justifying the use of HeLa cells to assay the effects of E2F4 overexpression.

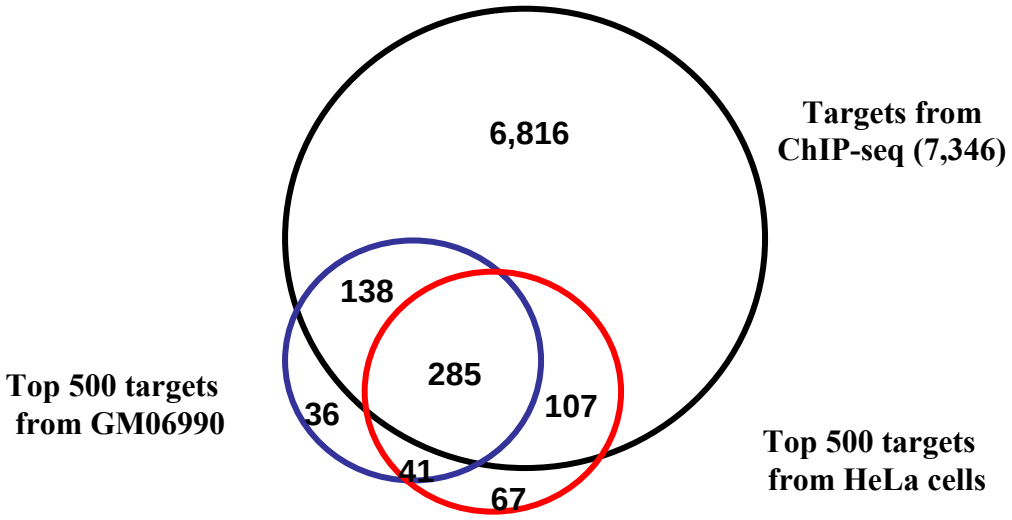

Figure 2-7. E2F4 target comparison between lymphoblastoid and HeLa cells.
Black, blue and red indicate numbers of targets from ChIP-seq with lymphoblastoid cells, numbers of targets from core promoter array with lymphoblastoid cells, and numbers of targets from core promoter array with HeLa cells, respectively. Targets from core promoter arrays were obtained from (Xu et al, 2007).

Transient overexpression of E2F4 alone did not trigger dramatic expression changes, even though some E2F4 responsive genes were up- or down- regulated. This result is likely due to low levels of its cofactors, which are required for E2F4 localization and binding. We therefore performed expression profiling after co-transfecting E2F4 with its cofactors (DP-1 and RBL2). Co-transfection resulted in increased levels of mRNA and protein for E2F4 and its cofactors (Fig. 2-8A and 2-8B).

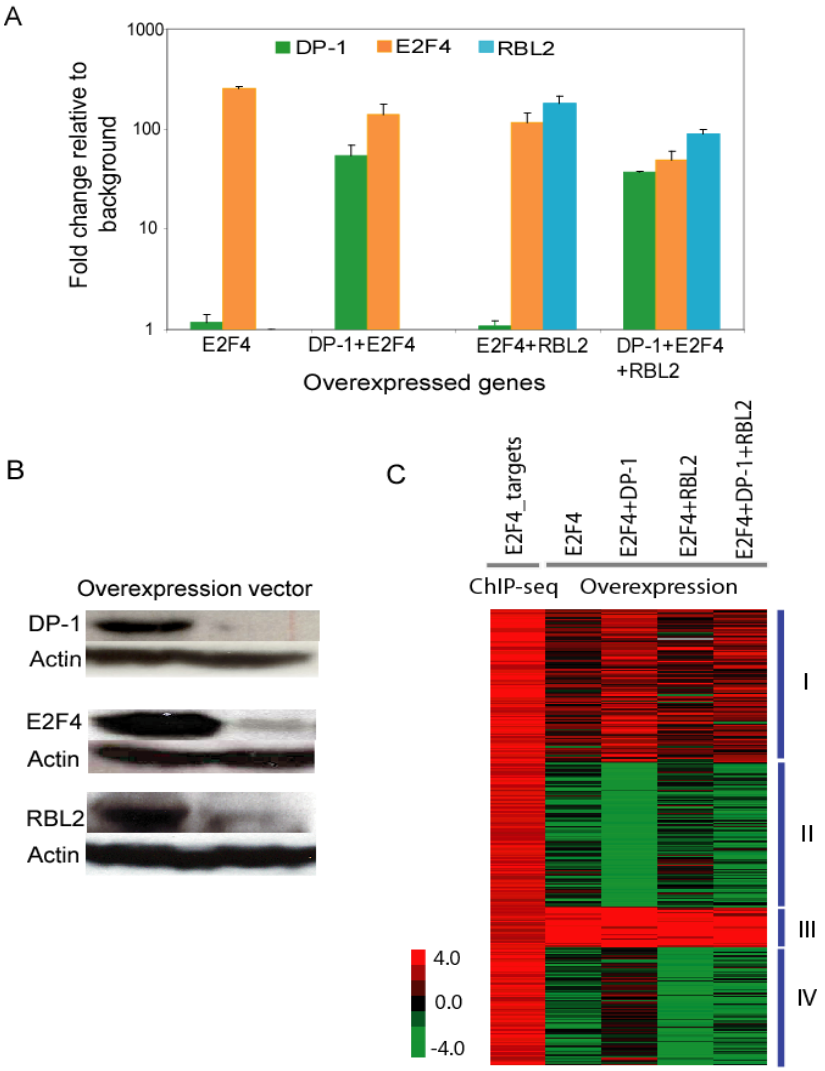

Figure 2-8. Overexpression of E2F4 and its cofactors (DP-1 and RBL2).
(A) Quantitative PCR verification of increase in mRNA of E2F4 and its cofactors. GAPDH was used as an internal control and the log-scaled Y-axis shows the fold increase of the indicated mRNA relative to the empty vector control. (B) Western blotting confirming overexpression of E2F4 and its cofactors at the protein level. Empty vector was used as a control. (C) K-means clustering of E2F4 targets identified by ChIP-seq along with gene expression data obtained in 4 different overexpression conditions. The expression data plotted is the expression value relative to that of a vehicle transfection control. The significance value (X) obtained from error model analysis was used for the clustering. A significance value of 3.3 corresponds to a P-value of 0.001. ChIP score was transformed to natural log.

We performed at least 2 biological replicates of the expression arrays and used an error-model to identify statistically significant genes whose expression was altered in response to overexpression of the regulators (Hu et al, 2007). Compared to overexpression of E2F4 alone, co-transfection of E2F4 and its cofactors increased the number of targets that showed significant expression changes (Table 2-6). Overall, combinatorial overexpression of E2F4 and RBL2 or E2F4 and DP-1 caused the downregulation of more E2F4 target genes than overexpression of E2F4 alone. However, co-expression of E2F4 and DP-1 also activated several genes.

Table 2-6. Number of up- or down- regulated genes after overexpression of E2F4 and its cofactors.

(p<0.001)	**Number of expression-changed genes**	
Overexpression	Up-regulated	Down-regulated
E2F4	167	128
E2F4+DP-1	314	341
E2F4+RBL	105	171
E2F4+DP-1+RBL2	228	281

P-value was calculated using an error model (Hu et al, 2007).

K-means clustering revealed four distinct clusters of genes whose expression was significantly altered: genes activated by overexpression of E2F4+DP-1 (cluster I), genes repressed by overexpression of E2F4+DP-1 (cluster II), genes activated by E2F4 alone (cluster III), and genes repressed by E2F4+RBL2 (cluster IV) (Fig. 2-8C). Cell cycle genes were highly enriched in both cluster I and cluster IV. Cluster I contained DNA replication and repair genes, while response to endogenous stimulus and programmed cell death were categories enriched in cluster III. These results indicate that E2F4 can function as either an activator or a repressor of transcription, and is involved in diverse physiological processes.

More importantly, several genes whose expression is positively associated with cell cycle progression were activated by E2F4 and its cofactors overexpression (CDC6, CDCA5, CEP55, MYBL2, RPA1, SGOL2, and SMC3). Only a subset of E2F4 binding targets showed significant expression changes, which suggests that the regulation of E2F4 target genes may be more complex than currently perceived. Interestingly, even the low scoring ChIP-seq binding targets were just as likely to be differentially expressed as the high-scoring ChIP-seq targets, and are therefore likely to be just as biologically meaningful (Figure 2-9).

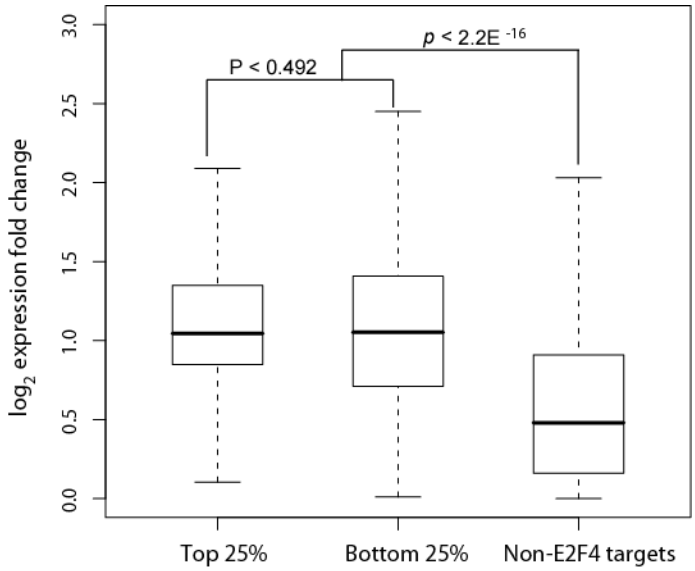

Figure 2-9. Box plot shows no expression difference between high ChIP-score E2F4 targets and low ChIP-score E2F4 targets.
Y-axis represents log2 expression fold change after E2F4 overexpression. High scores are from top 25% of E2F4 targets and low scores are from bottom 25% of E2F4 targets. Non-E2F4 targets represent genes that were not bound by E2F4 in our ChIP-seq. P-values were calculated by t-test.

E2F4 can regulate miRNAs

miRNAs have been implicated in fine tuning gene expression by cleaving target mRNAs or inhibiting their translation, and some of them cooperate to regulate specific cellular events (Croce, 2009; Stark et al, 2005; Sun et al, 2005). The expression of miRNAs is modulated by TFs and reciprocally, TFs are also regulated by miRNAs (Lal et al, 2009; O'Donnell et al, 2005; Sampson et al, 2007). To address whether E2F4 potentially regulates miRNAs, we first compared E2F4 binding sites with predicted human miRNA promoters that were identified based on histone modification signatures (Marson et al, 2008) and found 41 putative miRNA targets of E2F4 (Appendix D).

Since miRNA promoters are not well-defined, we also examined E2F4 binding sites located within 10 kb upstream of mature miRNA coding sequences. For this latter analysis, we excluded miRNAs present within exonic or intronic regions as it was not possible to assign E2F4 binding sites unambiguously to the miRNA or its parent gene. We thus identified an additional 161 miRNAs that showed E2F4 binding within 10 kb upstream of their coding regions (Appendix E). E2F4 showed strong binding to the putative promoters of the mir-17-92 cluster and let-7a, which are highly conserved miRNAs, as well as miR-22, an exonic miRNA (Fig. 2-10A).

Quantitative ChIP-PCR confirmed that E2F4 was indeed recruited to these three miRNA promoters in lymphoblastoid cells (Fig. 2-10B). To investigate whether E2F4, either by itself or in combination with its cofactors, could regulate these miRNAs, we first established that E2F4 did bind to the promoters of these miRNAs in HeLa cells also (Fig. 2-9B). We then used a quantitative TaqMan qPCR assay to measure expression changes of those three miRNAs in response to overexpression of E2F4 and its cofactors. All three miRNAs showed modest down-regulation upon overexpression of E2F4 alone or in combination with its cofactors (Fig. 2-10C).

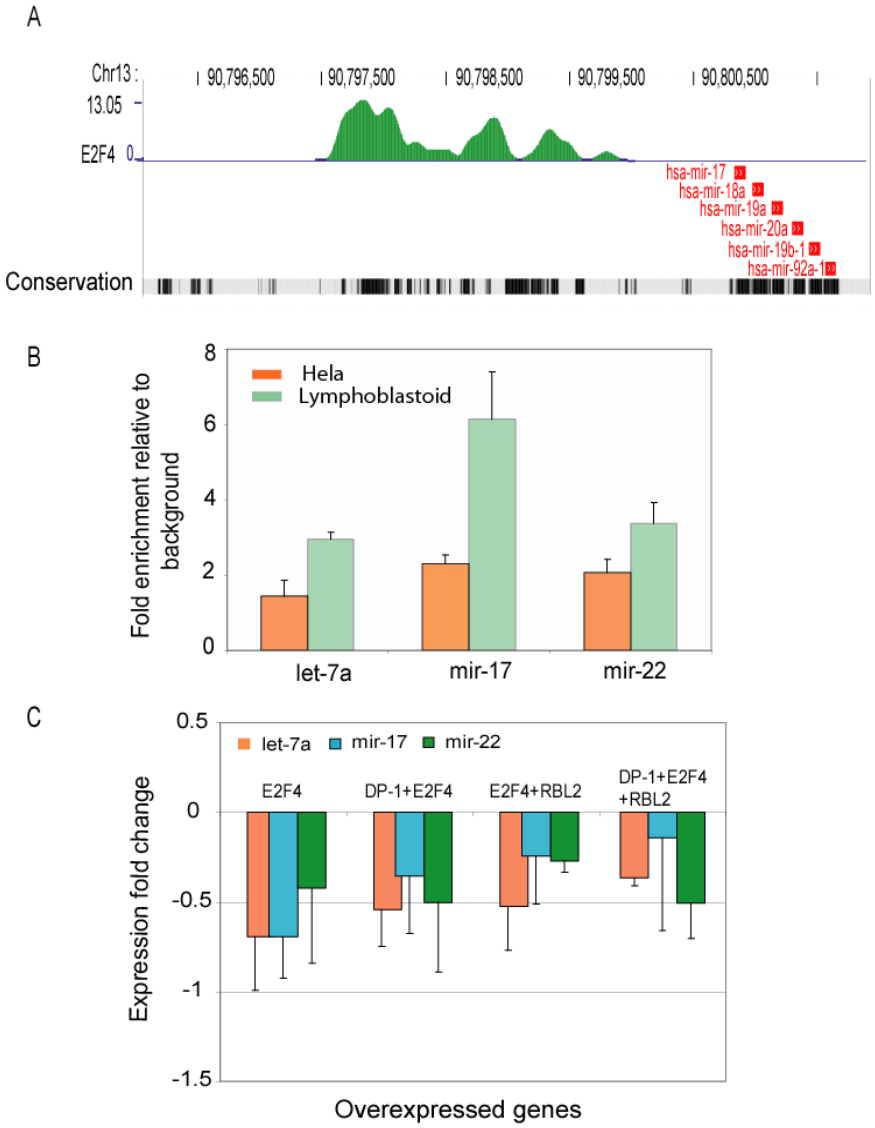

Figure 2-10. E2F4 can regulate miRNAs.
(A) ChIP-seq data showing E2F4 binding within 10 kb upstream of the mir-17-92 cluster. The positions of the miRNAs are shown in red. The bottom track shows phylogenetic conservation across vertebrates species (Vertebrate Multiz Alignment & PhastCons Conservation: http://genome.ucsc.edu) with darker vertical bars indicating greater conservation. (B) Quantitative PCR verification of E2F4 binding sites upstream of indicated miRNAs in lymphoblastoid and HeLa cells. (C) TaqMan qPCR data for miRNA expression upon overexpression of E2F4 and its cofactors. Different combinations of E2F4 overexpression with its cofactors caused a modest decrease in the expression of all three miRNAs. The data plotted is the $\log_2$ of the expression relative to RNU66 which served as the internal control.

2.4 DISCUSSION

The 16,246 E2F4 binding sites in the human genome that we identified by ChIP-seq are consistent with the number estimated by extrapolation from the sites previously identified using tiling microarrays covering 1% of the human genome (187 sites) (Xu et al, 2007)**.** About 56% of E2F4 sites were found at promoters, and the average binding profile of E2F4 relative to a gene showed a preference of E2F4 to bind near the TSS. This finding also agrees with previously published E2F4 ChIP-chip data (Xu et al, 2007). Overall, our E2F4 binding site analysis suggests that E2F4 mainly regulates the expression of target genes by being recruited to their core promoters. However, our unbiased ChIP-seq approach revealed a significant proportion of distal E2F4 sites that have not been noted before. This implies that in addition to regulating target genes by binding to the core promoter, E2F4 may be involved in additional modes of gene regulation.

Many TF binding sites in eukaryotes occur far away from TSSs and these distal regulatory regions are believed to have important physiological roles (Kim et al, 2007; Kimura-Yoshida et al, 2004)**.** For instance, a TF can play diverse roles by interacting with different cis-regulatory elements such as enhancers, insulators, or silencers. It is reasonable to speculate that some distal E2F4 binding sites function as enhancers or silencers to modulate target gene expression. Half of the distal E2F4 sites showed histone modification signatures characteristic of enhancers, suggesting that E2F4 may act like an enhancer at specific loci. Based on luciferase reporter gene assays we confirmed that

some of our distal E2F4 sites can function as enhancers. However, 5 out of 10 distal E2F4 binding sites did not show enhancer activity in the luciferase reporter assays even though those sites were highly enriched with enhancer marks of histone and p300 binding. This discrepancy may be in part because enhancers are cell-type specific, and the histone modification data were generated in a different cell type from the ChIP-seq data, and in part because histone modifications are not sufficient to fully specify enhancers. Nonetheless, our study suggests that in addition to a role at core promoters, E2F4 may act as a long-range regulator**.**

We found that E2F4 binding sites were highly enriched in bidirectional promoters, which is consistent with previously published data (Lin et al, 2007)**.** Bidirectional promoters may be an efficient way to modulate gene expression where the same DNA element regulates two different downstream genes at the same time**.** Genome-wide studies of bidirectional promoters in several mammalian genomes have suggested that they are evolutionarily conserved and functionally related in certain categories like DNA repair (Adachi & Lieber, 2002; Trinklein et al, 2004). We also found that E2F4 can modulate most E2F family members, as well as its own cofactors. Comparing our E2F4 target genes with all known RBL1 and RBL2 targets revealed that almost all cofactor targets overlapped with E2F4 targets. A number of studies have shown that E2F promoter specificity is determined by its cofactors. For instance, E2F4-p130/RBL2 is a major complex in quiescent cells, whereas an E2F4-pRB or –p170/RBL1 complex is important in the G_1 phase of the cell cycle, which suggests distinct roles of cofactors in different

cell cycle stages (Ikeda et al, 1996; van der Sman et al, 1999). The facts that E2F4 binds and may directly regulate its cofactors and family members suggest the possibility of feedback loops where the activity of E2F4 can be potentiated.

Motif analysis revealed that the canonical E2F4 motif was present in only 5% of the 16,246 E2F4 binding sites. This result implies the possibility that E2F4 uses other unknown motifs or is recruited to target promoters by the aid of other cofactors. *De novo* motif analysis using DRIM discovered 5 putative novel motifs. All motifs were found to be positioned near the peak of the ChIP-seq signal. As a corollary, this suggests that the peak position identified by our algorithm from ChIP-seq read data denotes the actual binding site of the protein. This level of resolution has not been achieved before in previous studies of E2F4 binding since they all used lower resolution tiling array approaches to identify binding sites.

Pathway analysis suggested that all E2F4 motifs were likely used to regulate the cell cycle pathway. Specifically, motif 1 (RTTYGAA) which was similar to a cell cycle repressor element, CHR (cell-cycle homology region; TTGAA) where E2F4/RB complexes were recruited (Yang et al, 2008; Zwicker et al, 1995), was highly enriched only among cell cycle pathway genes. Co-enrichment analysis identified several TFs that may co-regulate genes with E2F4. For example, many E2F4 targets such as E2F1, b-MYB, and HSORC1 contain SP1 motifs near E2F binding sites (Li et al, 1997); MYB and YY1 are known to be transcriptional partners of several E2F proteins (Giangrande et al, 2004; Schlisio et al, 2002; Zhu et al, 2004); the constitutively expressed factor, NF-Y,

binds to several cell cycle related E2F target promoters, and helps other regulatory proteins (PCAF and p300) gain access to target promoters to activate downstream genes (Caretti et al, 2003).

E2F4 was classified as a repressor of cell proliferation because it binds to its target promoters involved in cell cycle progression in G_0/G_1 and represses them. Even though E2F4 was previously known as a repressor, some studies introduced the possibility that E2F4 may function as an activator by showing that it was able to trigger cell proliferation. In addition, overexpression of E2F4 in transgenic mice induced cell propagation in the basal layer of the epidermis (Lukas et al, 1996; Pierce et al, 1998). Our genome-wide identification of E2F4 binding targets and transient perturbation of E2F4 followed by gene expression profiling indicate that E2F4 can indeed function as either an activator or a repressor of transcription. In particular, cluster I, consisting of genes activated by E2F4 and DP-1, contains genes implicated in positive regulation of the cell cycle, suggesting that E2F4 may function as a cell cycle activator.

Our data further revealed that E2F4 is capable of repressing the expression of several miRNAs such as the mir-17-92 cluster, mir-22, and let-7a, albeit by modest amounts. The mir-17-92 cluster, encoding 6 miRNAs, is known to be regulated by MYC, E2F1, and E2F3, and this regulation promotes cell proliferation (O'Donnell et al, 2005; Pickering et al, 2009; Woods et al, 2007). E2F4 may mediate its anti-proliferative role partly by repressing the mir-17-92 cluster. E2F4 is not only capable of repressing the expression of E2F1-E2F3, thus indirectly downregulating the mir-17-92 cluster, but also

binds to the mir-17-92 cluster promoter and directly regulates it, suggesting that it mediates a feedback loop for the regulation of the miR-17-92 cluster. Another miRNA target of E2F4, let-7a, is able to downregulate the expression of MYC as well as trigger cell cycle arrest (Sampson et al, 2007). Thus, E2F4 can not only regulate the expression of MYC directly (Chen et al, 2002), but also indirectly via let-7a, suggestive of another regulatory feedback loop. In summary, our genome-wide E2F4 target analysis reveals diverse functions of E2F4 and provides support for E2F4 functioning both as a long-range transcriptional regulators of mRNAs as well as a miRNA regulator, which allowed us to gain insights into understanding the versatile roles of this member of the E2F family.

Chapter 3: Lineage-specific and combinatorial usage revealed by genome-wide binding site studies of CTCF, MYC, and Pol II in multiple human cells

3.1 INTRODUCTION

In order to maintain cellular life, cells must rapidly and appropriately respond to various environmental stimuli by regulating the expression of a specific group of genes. Sequence-specific transcription factors (TFs) can recognize functional DNA elements in a sequence-specific manner, enabling cells to cope with diverse internal and external stimuli by regulating only a subset of their target genes (Vaquerizas et al, 2009). In addition, cellular diversity in a multicellular organism like the human is achieved in part by distinct programs of gene expression at the level of transcription, which in turn are mediated by sequence-transcription factors (TFs). Human has approximately 1,400 sequence-specific TFs (Vaquerizas et al, 2009). Identifying the genomic binding locations of TFs offers a means of understanding how their activities shape gene expression. Recent genome-wide studies have revealed more precise locations of functional elements including promoters, enhancers, insulators, and silencers with which several sequence-specific TFs interact (Cuddapah et al, 2009; Kim et al, 2007; Kim et al, 2010; Rada-Iglesias et al; Rada-Iglesias et al, 2010; Zill et al, 2010). However, this

location information is available for only a limited number of TFs in a few cell lines so that the majority of TFs still remain to be studied.

The Encyclopedia of DNA Elements (ENCODE) pilot project investigated 1 % of the human genome (30Mb) covering 44 genomic regions including protein-coding and non-coding loci (Birney et al, 2007). In its current second phase, the ENCODE Consortium is scaling up to whole genome studies in order to identify genome-scale cis-regulatory elements in a wide variety of cell lines. As a part of this ENCODE2 project we investigated the genome-wide binding sites of c-Myc (MYC), CTCF, and RNA polymerase II (Pol II) on the human genome.

In multiple signaling cascades, MYC plays an important role as a central hub integrating diverse internal and external stimuli (Wierstra & Alves, 2008). MYC, as a global regulator of transcription, can regulate approximately 15 % of human genes (Dang et al, 2006; Meyer & Penn, 2008) that are implicated in a wide range of biological functions including cell cycle progression, differentiation, apoptosis, DNA repair, angiogenesis, chromosomal instability, and ribosome biogenesis (Adhikary & Eilers, 2005; Dai & Lu, 2008; Dang, 1999; Knoepfler et al, 2006). In addition, MYC is a crucial factor for lineage-specific cell growth and metabolism so that it decides cell fate (Grandori et al, 2000). However, it is not clear how many lineage-specific genes are under regulation by MYC in the whole human genome.

The CCCTC binding factor (CTCF) is evolutionally highly conserved and ubiquitously expressed. CTCF contains 11 zinc-fingers DNA binding domains through a combinatorial use of which CTCF can be recruited onto diverse cis-regulatory sequences (Bell et al, 1999; Burcin et al, 1997; Filippova et al, 1996; Vostrov & Quitschke, 1997). This versatile binding capacity of CTCF allows multiple regulatory functions including gene activation as well as repression, hormone-responsive gene silencing, imprinting of genetic information, enhancer blocking, and chromatin insulation (Burcin et al, 1997; Filippova et al, 1996; Hark et al, 2000; Vostrov & Quitschke, 1997; Vostrov et al, 2002). Aberrant expression of either CTCF or MYC can cause detrimental consequences such as developmental disorders, disease, and a wide range of cancers (Ladomery & Dellaire, 2002; Ohlsson et al, 2001; van Riggelen et al, 2010).

Pol II is responsible for synthesizing precursors of mRNA, miRNA, and most snoRNA from DNA as a template (Sims et al, 2004). Following signal transduction cascades, activated sequence-specific TFs bind onto cis-regulatory elements of genes and recruit co-activators including histone modifying enzymes as well as chromatin remodelers. These sequential recruitments further facilitate open chromatin, enabling general transcription machineries including Pol II and several auxiliary factors to assemble near the transcription start sites (TSS) of genes. Most recent genome-wide mapping of Pol II in diverse tissues in mice identified many novel promoters and revealed cell-type specific alternative promoter usage (Sun et al, 2011).

Recent studies have also reported the interaction of Pol II with CTCF as well as MYC (Chernukhin et al, 2007; Rahl et al, 2010). CTCF has been shown to interact with Pol II both in vitro and in vivo, with significant co-localization of Pol II and CTCF observed in the nucleus (Barski et al, 2007; Chernukhin et al, 2007). In addition, CTCF-Pol II protein complexes are found in distal genomic regions, 1.5-15kb away from the nearest (TSS), and remain intact until Pol II release (Chernukhin et al, 2007). Genome-scale location analysis of CTCF has also revealed that approximately one-third of about 20,000 CTCF binding sites are located in protein-coding regions of the human genome (Barski et al, 2007). However, it is unclear how many Pol II binding sites are co-localized with CTCF genome-wide, whether there is lineage-specific co-localization in various tissue types, and how the interaction between Pol II and CTCF affects expression of their target genes. MYC can promote phosphorylation of the C-terminal domain of Pol II as well as mRNA cap methylation (Chernukhin et al, 2007). Most recently, c-MYC has been reported as a major regulator of the release of paused Pol II more so than recruiting Pol II at its promoters (Rahl et al, 2010) .

In order to identify the genome-wide binding sites of CTCF, MYC, and Pol II in diverse cell lines and elucidate possible combinatorial TF binding effects, we performed chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-seq) experiments. Across eleven cell lines we found an average number of binding sites on the order of 45,000 for CTCF, 30,000 for Pol II and 8,000 for MYC. Among those we discovered many cell-type specific binding sites that may render lineage-specific

properties. Interestingly, we also found that binding of either CTCF or MYC upstream of a gene, or even in the gene body, increases the expression of their targets genes, and that combinatory binding of MYC and Pol II notably enhanced expression of their target genes compared to binding of MYC or Pol II alone. Thus, our genome-scale investigation of sequence-specific TF binding sites advances the genome-wide understanding of the categories of genes governed by MYC, CTCF, and Pol II in diverse cell types, how the combinatory binding of Pol II with MYC or CTCF affects the expression of their target genes, and how and to what extent TFs are involved in linage-specific expression in diverse tissues.

3.2 MATERIALS AND METHODS

Cell culture

The ENCODE Consortium has designated GM12878, K562, HeLaS3, HepG2, HUVEC, NHEK and H1ESC cells as Tier 1 and Tier 2 cell lines. The source and cell growth conditions for these cells are described at the ENCODE web site (http://genome.ucsc.edu/ENCODE/cellTypes.html). Additional cell types analyzed in this study listed in Table 3-1 were cultured under standard culture conditions. Cells were grown to appropriate numbers and processed for chromatin immunoprecipitation.

ChIP sequencing

ChIP assays were performed as described previously (Lee et al, 2011). Briefly, cells were cross-linked with 1% formaldehyde for 10 min at room temperature. The cross-linked cells were sheared by sonication until average fragmented DNA size reached 500 bp, then TF-DNA complexes pulled down with specific antibody for CTCF (07-729, Millipore), MYC (SC-764X, Santa Cruz Biotech), and Pol II (MMS-126R, Covance Inc.). ChIP for Pol II was performed by Ryan, a graduate student in Iyer lab and ChIPed DNA was sequenced primarily using single-end Illumina Solexa sequencing technology, with one replicate from SOLID sequencing.

Peak calling and statistical correction

Sequencing generated 32-36 bp of short reads from the ends of ChIP-enriched DNA fragments and from corresponding non-enriched control DNA fragments (Input). These short reads were mapped back to the human genome (hg18) using the Maq aligner (Li et al, 2008). Total number of mapped reads from ChIP and Input sequencing are listed in table S1. In order to identify precise binding sites of a TF from high-throughput sequencing data, we used a Parzen window based algorithm as described previously with minor modifications (Lee et al, 2011; Shivaswamy et al, 2008). Each aligned read was assigned a value representing the frequency of observing that read in the sequencing library. After extending reads in the 3' direction by half (67 bp) the sequencing library fragment length, a Gaussian kernel with a defined bandwidth was applied to weight the occupancy scores based on the proximity of neighboring nucleotides. Local maxima of these Parzen scores were used to define binding sites along with IQR-based calculations to determine the adjacent region of highest read density. The total number of reads was recorded for each binding site, along with chromosomal locus information and the position of maximum weighted read density. The resulting set of candidate binding sites was then subjected to input correction, filtering, and statistical significance determination steps.

First, to normalize for non-specific binding represented by the Input control for each cell line (input correction), each binding site was paired with the corresponding

Input site (within 200bp) with the highest read count. A binomial P-value was computed for each binding site under the null hypothesis that ChIP and Input reads are equally likely. A simple ratio of total ChIP to Input reads was used to normalize for differences in read depth before applying the binomial cdf; however, to avoid inflating read numbers, the library with higher sequencing depth was always scaled downward. Next, the binomial P-value was used to adjust the binding site's read count by calculating the number of ChIP reads there would be if no input were present, solving pbinom(input, chip+input, .5) = pbinom(0, corChip, .5) for corChip. This binomial P-value corrected number of reads (binCorRd) score was recorded for each binding site and was used in all further occupancy score-based analyses.

Initial filters were then applied. Input-dominated binding sites were discarded and only sites where the sequencing-depth-scaled ChIP read count exceeded Input were retained. Binding sites in the standard Duke-defined ENCODE2 filtering regions were also discarded. Finally, sets of high-confidence sites were identified based on the ECDF of the filtered binding sites. Strictly quantitative determination of target cutoff levels is an elusive goal given the large number of data sets under analysis, their temporal separation as underlying sequencing technology advanced, differences in both library preparation and sequencing-run quality, and notable differences in binding characteristics among the factors. Qualitatively, CTCF binding sites tend to be narrow and sharp while the dynamic Pol II transcription machinery produces broad, dense binding signals, and MYC sites seem to be less prevalent than either of the other two. As a result, a different top

percentage level of highest-scoring binding sites was chosen for each TF: 4% for CTCF, 2% for Pol II and 0.5% for MYC, and a corresponding threshold binCorRd score cutoff was identified for each data set. Since this binCorRd score has both read density and significance probability components, minimum and maximum score considerations were applied across all datasets to account for experiment-specific quality differences. Data sets with target percentile cutoff scores below that roughly corresponding to a binomial P-value of .0005 had their thresholds adjusted upwards (removing the lowest scoring sites) and data sets with scores corresponding to a binomial P-value of 1E-10 had thresholds adjusted downward (capturing additional high-scoring sites). The final count of significant binding sites identified for each data set, along with the corresponding cutoff scores and percentage of top-scoring sites represented, is shown in Table 3-2.

Mapping binding sites to gene features

CTCF, MYC, Pol II binding sites were mapped to within ± 2 kb from the TSSs of all annotated genes generated by combining gene lists from RefSeq, UCSC, Ensemble, Vega, and SIB downloaded from UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgTables?command=start), resulting in a total of 348,592 TSSs. We defined a distinct gene as a combination of chromosome number, start, and number of exons unique across annotation files. To assess the number of sites mapped to different genomic features, we assigned one site to only one feature using following hierarchy: promoter > upstream > intron > exon > intergenic, since TFs in general have a preference for binding near the

TSSs of genes. A promoter was defined as 2 kb up- and downstream from the TSS, upstream as between 2 kb and 20 kb upstream from the TSS. Binding sites that could not be mapped to within 20 kb upstream of any TSS, nor to any exon or intron, were termed intergenic. Genes that had TF binding sites within the promoter were defined as TF target genes.

Mapping binding sites to CpG island

CTCF, MYC and Pol II binding sites in CpG islands were investigated by mapping their binding sites in CpG islands downloaded from UCSC genome brower. Binding sites of a TF were assigned into CpG binding peaks as long as the sites were located within CpG islands.

Mapping binding sites of bidirectional promoters

We defined a bidirectional promoter as a genomic region that is not only exclusively upstream from the gene but also between the TSSs of two genes which were separated by maximum of 2kb in length and divergently transcribed from opposite strands. Based on this criterion, we identified 1233 bidirectional promoters corresponding to 2466 genes (11.06 %) among the 22,279 genes annotated in RefSeq. In order to investigate enrichment of CTCF, MYC and Pol II binding in bidirectional promoters, we, first, searched target genes of CTCF, MYC and Pol II by mapping their binding sites within 2Kb upstream from all TSSs annotated in RefSeq. Among those target genes, we examined number of genes which are regulated by bidirectional promoters. Using

hypergeometric distribution, we finally calculated significance of CTCF, MYC and Pol II enrichment in bidirectional promoters.

Profiling TSS/TTS binding

A 10 kb region around the TSS including both 10 kb upstream and 10 kb downstream was binned with a 50 bp bin, and then CTCF, MYC, and Pol II binding sites were mapped to each bin. The score of a peak that was mapped to each bin was assigned into the bin. A profile of average peak score across TSSs was generated by averaging corresponding bin scores across all genes. The same procedure was applied to generate TTS profiles.

Overlap analysis

Overlapping binding sites among different ChIP-seq datasets were evaluated by scanning overlapping sites across the genome whose centers resided within a 300bp window of each other. In this evaluation, we compared the binding sites of one ChIP-seq dataset at a threshold cut-off with the binding sites of another ChIP-seq dataset at a looser threshold obtained by multiplying the original threshold by 0.5 to avoid false positive cell-type specific binding sites arising to overlapping binding sites in another cell type that just missed the threshold.

Pol II analysis

Relationship between a Pol II binding pattern and its consequent gene expression was investigated in 4 Pol II binding groups (HH: high occupancy in both the promoter and the body of a gene; HL: high occupancy only in the promoter; LH: high occupancy only in the gene body; LL: low occupancy signal in both promoter and gene body) that were divided based on the occupancy signal intensity from proximal promoters (ranging from 2 kb upstream to 300bp downstream from a TSS) to gene bodies. First, we assigned occupancy score into promoters and gene bodies by mapping Pol II binding in them. Next, using empirical cumulative distribution function (ECDF), we rank both promoter and gene body scores from highest to lowest occupancy signal. To classify 4 Pol II binding patterns we used high occupancy threshold as 0.7 and low as 0.3 in ECDF. We considered HL group as genes having paused Pol II. We also determined the significance of enrichment of CTCF and MYC around Pol II sites in promoters as well as gene bodies using the hypergeometric distribution function.

Expression profiling

RNA expression profiling was carried out independently as part of the cell line phenotyping component of the ENCODE project. For GM12878, K562, HeLaS3, HepG2 and H1ES, RNA was generated from the same culture of cells used for ChIP; for HUVEC, NHEK, MCF7, FB8470, FB0167P and H54 cells, different cultures were used for RNA and for ChIP. Details of RNA expression have been previously described, and

these data were deposited to the Gene Expression Omnibus (GSE15805) (Boyle et al, 2011).

Motif analysis

Motifs of sequence-specific TF (CTCF and MYC) binding sites were investigated using discriminating matrix enumerator (DME) (Smith et al, 2005). First, we divided the binding sites of a TF into three group: strong (top 25%), moderate (middle 50%), and weak (bottom 25%) binding sites based on ChIP score, and considered the top 500 sites from each group for motif search. A 200 bp region obtained from 100 bp up-and down-stream sequences from the center of a peak was extracted from the human genome assembly hg18, and then applied for motif discovery. A random background was generated by sampling 200 bp of 100,000 sequences from the genome. Since approximately 15% of CTCF sites and 50% of MYC sites across the cell lines analyzed occurred in promoters ranging from 2 kb upstream to 2 kb downstream of TSSs of genes, this promoter binding portion of each factor was maintained in the random sample corresponding to each factor.

3.3 RESULTS

ChIP-seq identifies genome-wide high confidence binding sites for CTCF, MYC, and Pol II

We performed ChIP-seq for CTCF, MYC, and Pol II in 11 different human cell types including primary, disease, and cancer cells to identify their the genome-wide binding locations. In parallel with sequencing chromatin immunoprecipitated DNA we also sequenced a non-enriched input DNA control for each cell type. We generated at least two biological replicates of ChIP-seq data for each factor in all cell types except H1 ES and NHEK cells (Table 3-1).

Table 3-1. The replicates and aligned reads for all ChIP-seq as well as input data.

	Input		CTCF		MYC		Pol II	
	Replicates	# Aligned Reads	Replicates	# Aligned Reads	Replicates	# Aligned Reads	Replicates	# Aligned Reads
FB0167P	1	31,510,516	2	46,874,291	2	25,230,535	2	49,082,997
FB8470	1	40,979,563	2	39,266,345	2	10,955,174		
GM12878	1	16,568,390	3	30,830,198	2	27,762,943	2	41,965,165
H1ESC	2	21,110,780	1	14,563,540	1	18,719,974	1	22,033,234
H54	1	51,821,256	2	104,978,274	2	58,004,041	2	36,683,532
HelaS3	1	11,798,703	2	28,322,770	2	14,228,304	2	18,851,779
HepG2	1	11,929,226	2	14,063,412	3	25,203,575	4	36,930,350
HUVEC	1	14,898,732	2	22,204,482	2	30,953,678	2	28,353,668
K562	1	16,493,115	3	27,270,533	3	34,160,398	2	40,791,878
MCF7	1	29,404,931	2	45,850,125	2	80,691,282	2	58,363,979
NHEK	1	26,350,143	2	26,316,271	1	13,234,802	1	20,058,797
Average	24,805,941		36,412,749		30,831,337		35,311,538	
Total	241,354,839		353,665,950		313,914,171		304,032,382	
			Average		31,101,727			
			Total		1,212,967,342			

To identify binding sites of CTCF, MYC and Pol II, we first aligned the raw sequence reads onto the human reference genome, and then used a Parzen window kernel density estimation algorithm to detect binding sites from each replicate dataset (Methods). Next, we investigated the reproducibility of our ChIP-seq data by analyzing the overlap between two biological replicates for each factor in a given cell line. In most cell lines, CTCF and Pol II showed high consistency between replicates (80% overlap), and MYC exhibited moderate reproducibility ranging from 50-80% (Fig 3-1A). Therefore, we combined reads from all replicates from each factor-cell line combination, and then performed peak detection again to generate an initial set of candidate binding sites for CTCF, MYC, and Pol II (Fig.3-1B). In order to minimize false positives, we normalized TF binding scores with corresponding input scores after correcting for sequencing depth. We calculated *P*-values, normalized scores and determined appropriate thresholds for targets (Methods; Table 3-2). Most subsequent analysis was performed using putative binding sites at this target cutoff level, except where indicated.

Although there is some variation in the number of binding sites of each factors in 11 cell types, CTCF, MYC, and Pol II had approximately 45,000, 8000, and 30,000 sites on average, respectively. Particularly for Pol II, these numbers include partially overlapping sites in each cell and thus don't directly reflect the number of genes that are potentially targeted. For each factor, additional cell line continued to show additional binding sites rather than reaching saturation (Fig. 3-1C), implying many cell-type specific binding sites exist in diverse cell types.

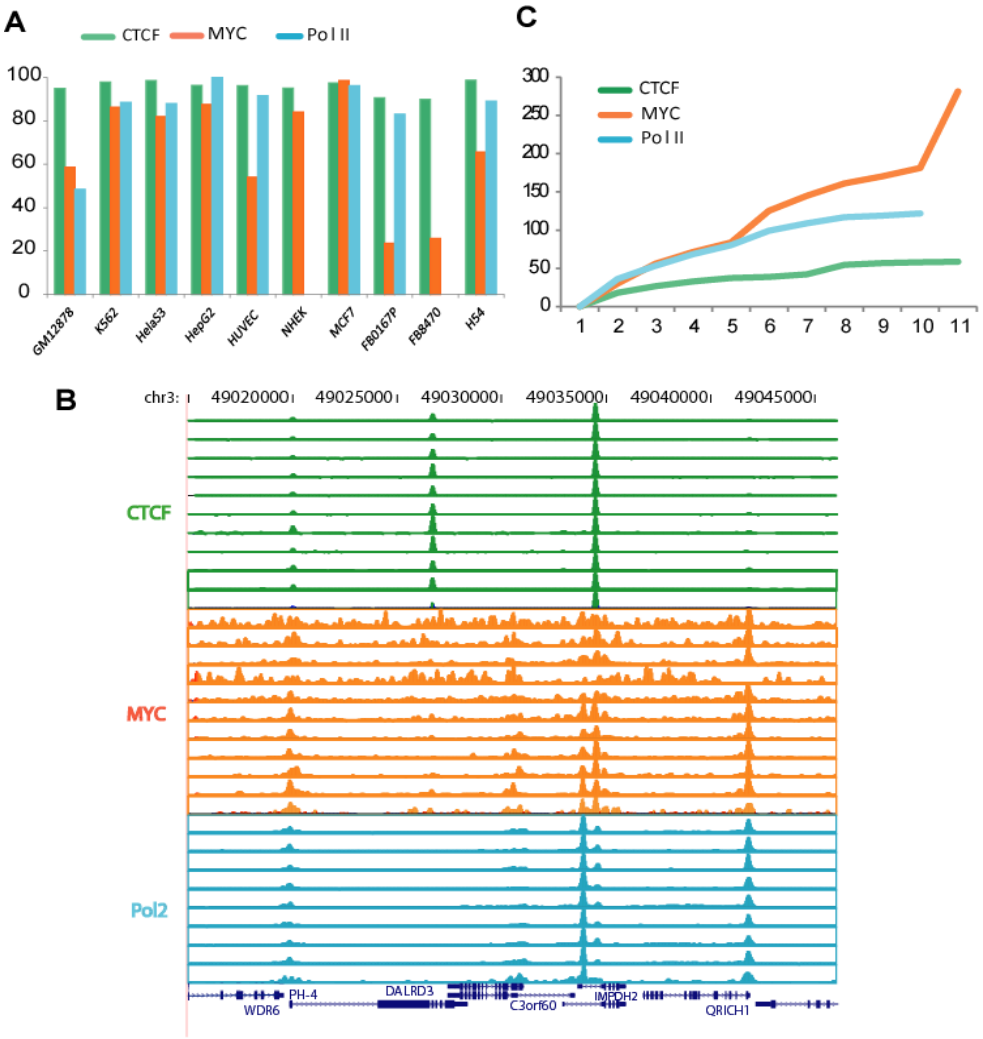

Figure 3-1. ChIP-seq produces genome-wide high confidence binding sites of CTCF, MYC, and Pol II in diverse cell lines.

(A) Overlap analysis between two biological replicates of ChIP-seq data for each factor in each cell line. The top 50,000 CTCF, 15,000 MYC, and 30,000 Pol2 binding sites from one replicate were compared with the top 65,000, 30,000, and 50,000 binding sites from the other replicate, respectively. The Y-axis shows percent overlap between replicates for the cell types indicated on the X-axis. Overlap values for Pol II are not shown for Keratinocytes and FB8470 because only one replicate was used for Keratinocytes, and no data was generated for Pol II in FB8470. (B) Occupancy signal in ChIP-seq tracks shows distinct TF binding sites relative to background. Blue, red, and green colors indicate CTCF, MYC, and Pol II occupancy signals respectively. Chromosome coordinate are shown on top. Each lane shows a TF ChIP-seq track from one cell line. Gene name and location were shown in bottom track, with direction of transcription with arrows. (C) Additional cell reveal increasing numbers of CTCF, MYC, and Pol II binding sites in the human genome.

Table 3-2. The Number of CTCF, MYC, and Pol II binding sites at a cutoff threshold in diverse cell lines

	CTCF			Pol II			MYC		
	Top %	Score	# Sites	Top %	Score	# Sites	Top %	Score	# Sites
FB0167P	3.5	10.97	46,424	2	18.46	36,325	0.5	17.72	7,100
FB8470	4.2	33.22	35,652				0.5	12.65	4,661
GM12878	3.7	10.97	50,366	2	20.06	31,118	0.39	10.97	8,134
H1ESC	4	21.25	40,890	2	25.94	25,089	0.5	19.36	7,274
H54	4	29.68	43,501	2	20.78	23,069	0.26	10.97	3,938
HelaS3	3.8	10.97	53,731	2	16.15	29,249	0.48	10.97	6,949
HepG2	4	16.66	43,234	2	13.64	42,550	0.5	13.14	11,623
HUVEC	3.6	10.97	45,175	2	21.82	33,924	0.5	12.82	11,862
K562	4	30.54	37,289	2	28.36	31,560	0.5	19.91	9,685
MCF7	4	22.98	45,705	2	12.73	41,224	0.92	33.22	15,435
NHEK	7.2	33.22	38,009	2	26.28	13,511	0.5	16.85	4,697
Average			43,355			30,762			8,305
Total			479,976			307,619			91,358

CTCF prefers to bind onto intergenic regions while MYC and Pol II mainly associate with promoters

Since it is widely accepted that sequence-specific TFs have a strong binding preference for the TSSs of genes (Ren et al, 2002; Tabach et al, 2007), we examined the average binding profile of CTCF, MYC, and Pol II around the TSS in 11 cell types. Consistent with previous studies, strong occupancy signal of binding sites was enriched around the TSS (Fig. 3-2A). Importantly, our ChIP-seq score normalization (Methods) allowed us to quantitatively compare occupancy scores and binding profiles across

factors and cell types. As Pol II binding sites are expected to show higher enrichment in the promoters or genic regions of transcribed genes rather than the intergenic regions, we found that Pol II, regardless of cell line, showed the strongest propensity to bind at TSSs. The two sequence-specific TFs (CTCF and MYC) also showed a strong preference for binding near TSSs (Fig. 3-2A). These three factors often occupied at TSS in a combinatorial manner, as illustrated in the example track image in Fig 3-1B. In contrast to the TSS profiles, all three factors showed depleted binding near the transcription termination sites (TTSs) of genes where Pol II disassembles its transcriptional machineries (Fig. 3-2B).

In addition to averaging TSS profiling, we also examined gene-wise occupancy signals of these three factors within ± 10 kb from TSSs. Strikingly, the patterns of occupancy signals were maintained across the cell types we analyzed for all three factors (Fig. 3-2C). Furthermore, subset of genes has its own unique occupancy pattern, in particular distinctive CTCF binding signal in distal as well as gene bodies of many genes and strong Pol II occupancy signal observed across the gene bodies of many genes.

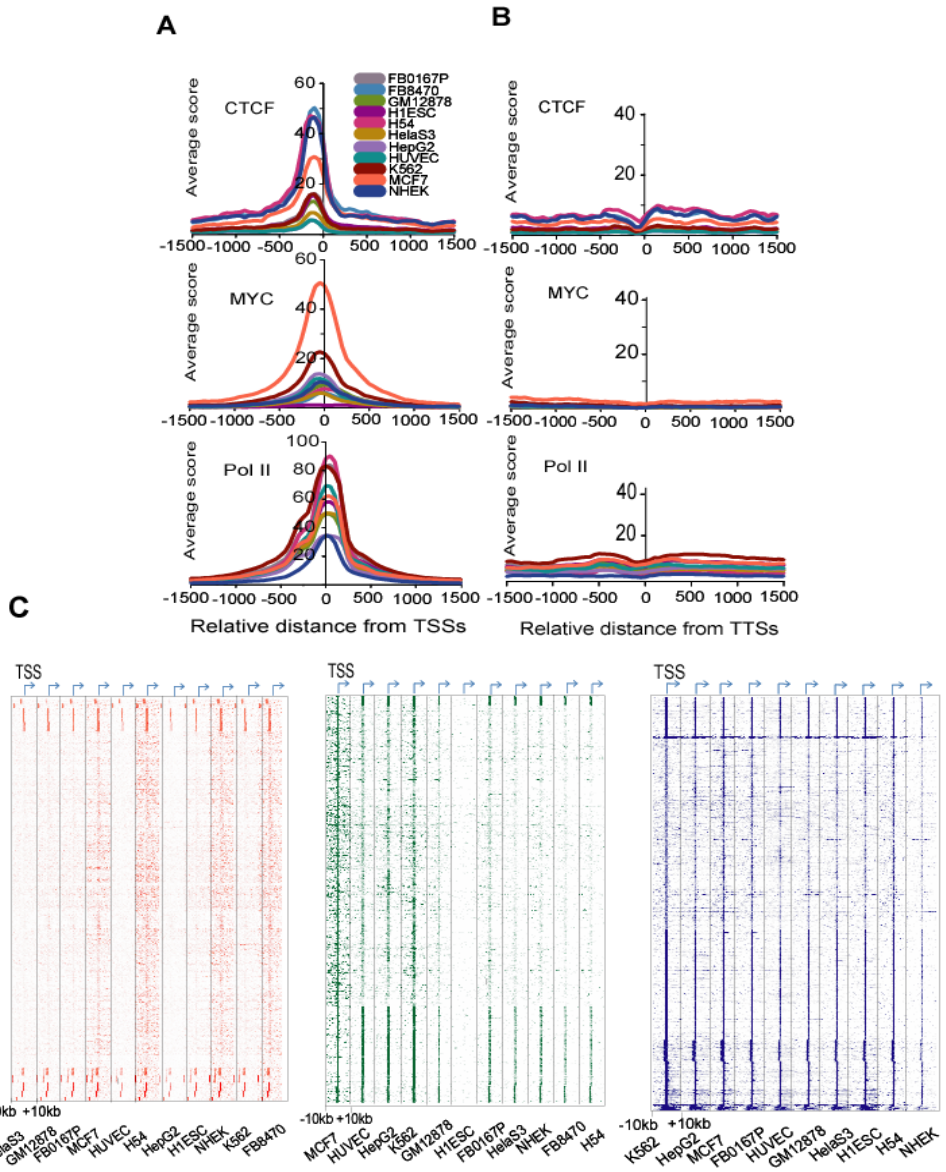

Figure 3-2. The genome-wide distribution patterns of CTCF, MYC, and Pol II binding sites in diverse cell types.

(A) Average binding profiles of CTCF, MYC, and Pol II sites within ± 1.5 kb from all annotated TSSs and (B) TTSs in 10-11 different cell types. Zero in X-axis indicates transcription start sites of all annotated genes. (C) Patterns of CTCF, MYC and Pol II occupancy around genes. The heat map shows normalized occupancy scores for each gene with available data (rows) within ±10 kb from TSS (columns are 100 bp bins). Arrow indicates TSS and direction of transcription. For each factor, data in the first listed cell line was clustered using K-means clustering and data for the other cell lines are displayed in the same order. The ordering of genes is not the same for the 3 factors. The number of genes listed in the Y-axis is 13,610 for CTCF, 10,440 for MYC, and 13,000 for Pol II.

Next, we investigated the distribution of CTCF, MYC, and Pol II binding sites across the genome by mapping their binding sites relative to RefSeq annotated genes or a combined gene set including RefSeq, UCSC, Ensembl and Vega annotated genes from the UCSC genome browser (http://genome.ucsc.edu/). We defined a region within ± 2 kb from a TSS as a promoter, between 2 kb and 20 kb from a TSS as an upstream region, and more than 20 kb away from a TSS (not including exons or introns) as an intergenic region. We found that each factor had a distinctive pattern in the distribution of its binding sites across the genome. For instance, MYC varied substantially, depending on the cell line, in the proportion of its binding sites at promoters from ranging from 45 % to 75 % (Fig. 3-3A). The ES cells however were an outlier in this regard, with only 6% of its binding sites at promoters. Unlike MYC, Pol II and CTCF showed a relatively constant proportion of their sites (~70 % and 15% respectively) at promoters across all the cell types we analyzed (Fig. 3-3A). More than 60% of CTCF binding sites, across all cell types, were in distal upstream and intergenic regions. This distinct distribution of CTCF binding sites from other factors is consistent with previous genome-wide binding studies of CTCF in individual cell types, which showed that CTCF binding sites were widely dispersed across the genome (Barski et al, 2007; Schmidt et al, 2010). Since TF binding to a proximal promoter is most likely to regulate the expression of its immediate downstream gene, comparing the proportion of TF binding sites in TSS to the proportion of all TSS bound by the TF provides an indication of promoter selectivity of a TF. This analysis showed that promoter binding at ~70 % of Pol II sites may regulate up to a half

of all genes and ~15 % of CTCF sites may regulate as much as a quarter of all genes in the genome (Fig. 3-3B). Both CTCF and Pol II showed small variations in the number of putative target genes across the different cell types, while MYC generally exhibited substantial variation, from 10% to 35% of genes, CTCF regulates functionally conserved genes across different cell lines, whereas MYC may modulate various functional classes of genes in different cell lines.

In addition to the promoter binding preference of MYC and Pol II, they showed significantly higher occupancy scores at promoters than in other genomic regions (Fig. 3-2F). In contrast to MYC and Pol II, CTCF showed significantly higher occupancy signals in upstream regions rather than promoters (Fig. 3-3C), consistent with a prominent role for CTCF as an insulator binding protein, functioning between a promoter and an enhancer. Consistent with these findings, about half of all MYC and Pol II sites were located in CpG-islands, which are known to be associated with promoters, whereas the majority of CTCF binding sites were located in CpG-depleted regions (Fig 3-4A). Conversely, the binding sites of these three factors in CpG islands were enriched for promoters (Fig 3-4B). Moreover, the binding sites of CTCF, MYC and Pol II in CpG islands had significantly higher occupancy scores than their sites in non-CpG-containing sites across all cell types (Fig. 3-4C). Taken together, these results indicate that MYC and Pol II regulate genes primarily by binding to proximal promoters, with MYC exhibiting greater diversity of gene targets across cell types, whereas CTCF modulates gene

expression of a more conserved set of targets by associating with distal cis-regulatory elements on the genome.

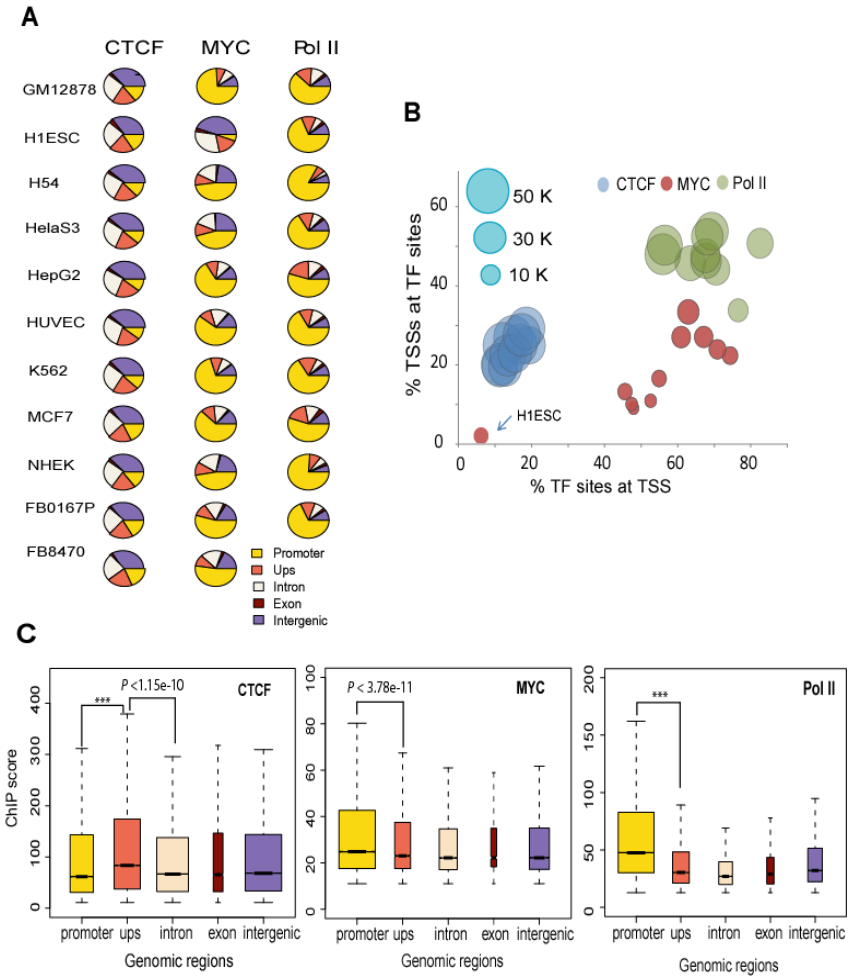

Figure 3-3. The genome-wide distribution patterns of CTCF, MYC, and Pol II binding sites in 5 different genomic regions in diverse cell types.

(A) Pie charts show the distribution of CTCF, MYC, or Pol II binding sites in 5 different genomic regions. A promoter is defined as a region within ± 2 kb from the TSS of a gene, upstream is between 2 kb and 20 kb upstream from the TSS, and intergenic is a region excluding a promoter, upstream, intron and exon. Each locus was coded with a different color. (B) Percent binding sites of each TF within ± 2 kb from TSSs and percent TSSs within ± 2 kb from binding sites of TFs. Circle size correlated with number of binding sites. Each color represents a specific factor. An arrow points a specific cell line. (C) Box plots showing ChIP-seq score distribution of CTCF, MYC, and Pol II binding sites in 5 different genomic regions across the cell lines we analyzed. MYC and Pol II binding sites in promoters had significantly higher ChIP scores than other genomic regions, whereas CTCF had higher ChIP score in upstream. *P*-value was calculated by Wilcoxon rank sum test. Three stars (***) indicate *P*-value of zero.

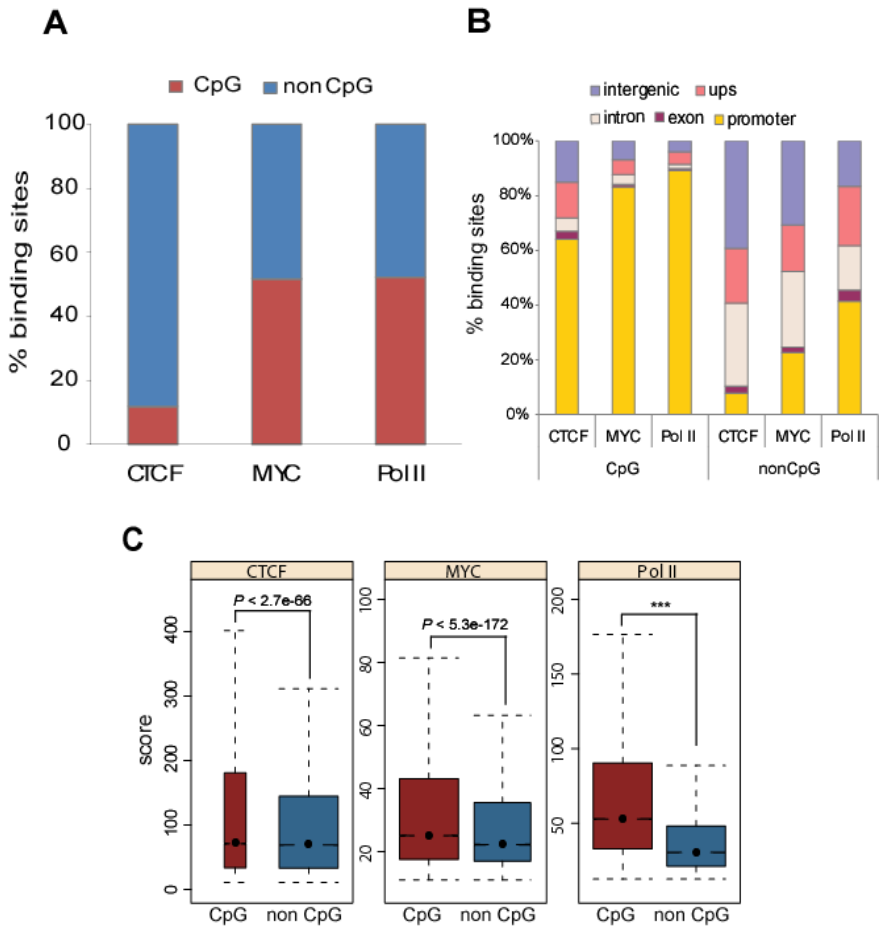

Figure 3-4. The genome-wide distribution patterns of CTCF, MYC, and Pol II binding sites in CpG and non-CpG sites.

(A) CTCF prefers to bind in non-CpG islands. X-axis shows a factor and Y-axis represents % binding sites in CpG and non-CpG loci. CpG and non-CpG sites are labeled with red and blue. (B) The distribution of CpG and non-CpG binding sites of CTCF, MYC, and Pol II in 5 different genomic regions. (C) Boxplots show a significantly higher occupancy signal of MYC and Pol II binding sites in CpG island than in non-CpG one. *P*-values were calculated by Wilcoxon rank sum test. Three stars (***) indicate *P*-value of zero.

CTCF, MYC and Pol II sites are positively correlated with gene density across the genome

The human genome has different gene density from one locus to another, with gene-rich or gene-poor regions are disseminated across the genome. (Lander et al, 2001). Since all three TFs occupied distal binding sites in addition to promoters, we investigated the relationship between TF binding sites and gene density. The correlation coefficient of CTCF, MYC, and Pol II binding sites with gene density in 2 Mb bins in K562 was 0.81, 0.79 and 0.82, respectively (Fig. 3-5A) and similar correlation was observed across all analyzed cell types (Fig3-5B), indicating a positive correlation between TF binding sites and gene-density. Interestingly, although our genome-wide binding analysis of CTCF showed a clear preference of CTCF for intergenic regions (Fig 3-3A), its binding was nonetheless positively correlated with gene density. Excluding CTCF binding sites within genes as well as up to 20 kb upstream of TSS did not completely abrogate the correlation between binding sites and gene-density (Fig 3-5B), suggesting that CTCF regulates gene expression not in the proximal location of genes, but at least in gene-dense neighborhoods.

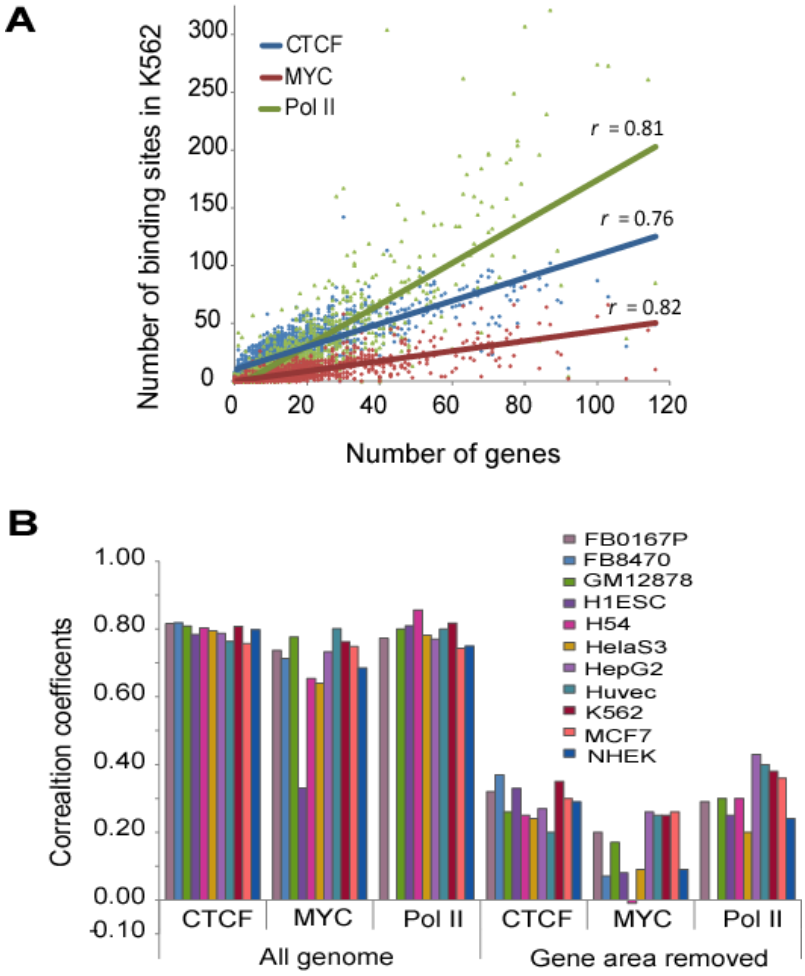

Figure 3-5. TFs' binding sites are positively correlated with gene density.
(A) Pearson correlation coefficient (*r*) was calculated using the number of a TF binding sites and the number of genes in 2Mb bin across the genome. Red, blue, and green represent CTCF, MYC, and Pol II respectively. A linear line was drawn by fitting data into linear regression. (B) Pearson correlation coefficients between the number of CTCF, MYC, and Pol II binding sites and gene density. Removing all binding sites of CTCF, MYC, and Pol II still shows a positive correlation. Shuffling of data exhibited no correlations(data not shown).

MYC is enriched in divergent promoters

It has been reported that the motifs of some sequence-specific TFs including GABPA, MYC, E2F1, E2F4, and YY1 are overrepresented in bidirectional promoters (Lin et al, 2007), and we have previously reported that E2F4 binding sites are overrepresented in bidirectional promoters (Lee et al, 2011). We therefore examined whether any of the transcription factors showed a bias in binding to bidirectional promoters. Based on annotations for 22,279 genes in RefSeq, there are 1233 promoters corresponding to 2466 bidirectionally transcribed genes in the human genome. In the majority of cell types, bidirectionally transcribed genes were significantly overrepresented among the targets genes of Myc where it bound within 2 kb of the TSS (Fig. 3-6A). This overrepresentation of bidirectional promoters was specific to Myc binding sites as it was not observed for CTCF and Pol II binding sites. We also found that in most cases, the binding of CTCF, MYC, or Pol II at bidirectional promoters activated both genes equally, regardless of the distance of its binding site from a TSS (Fig 3-6B).

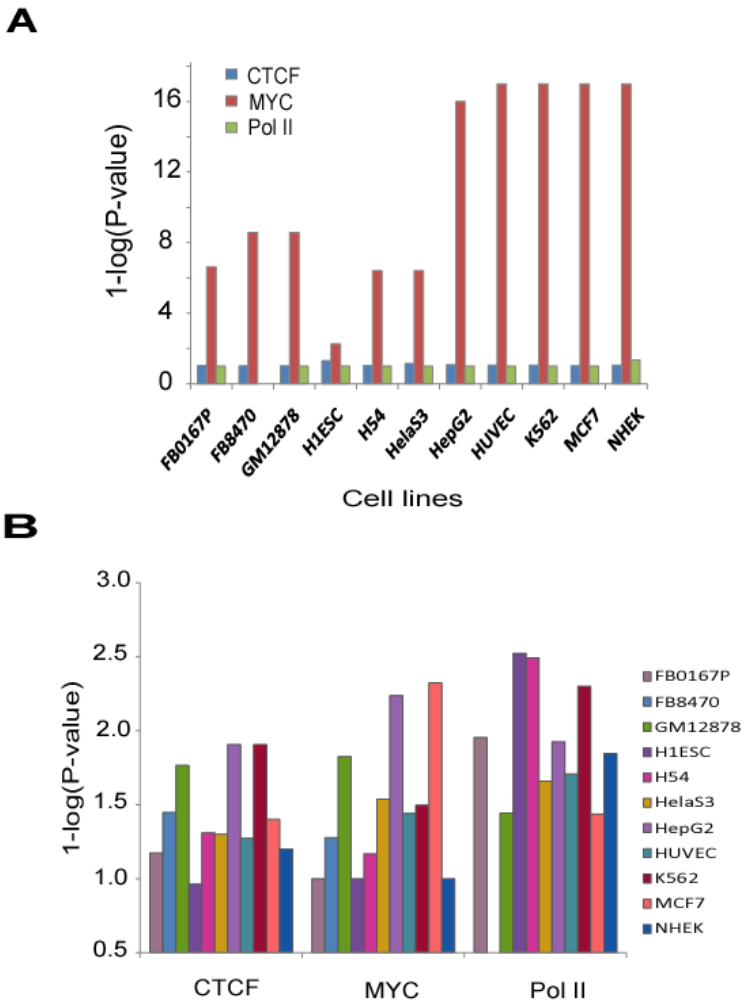

Figure 3-6. MYC is enriched in bidirectional promoters.
(A) Bar graphs show MYC enrichment in bidirectional promoters. X-axis shows different cell types. Y-axis shows 1-log (*P*-value) calculated using hypergeometric distribution. (B) TFs binding in bidirectional promoters activates both genes equally, regardless of its binding distance from transcription start site of genes. Value 3 in Y-axis indicates *P*-value of 0.01.

CTCF and Pol II sites are ubiquitous, whereas MYC sites are cell-type specific

Different tissues require distinct expression patterns of certain groups of genes to fulfill cell-type specific demands or to determine cell fate during differentiation. Visual inspection of CTCF, MYC and Pol II sites in the genome browser showed that all three factors had cell-type specificity to some extent (Fig. 3-7A). To examine the cell-type specific binding sites of CTCF, MYC, and Pol II, we first analyzed overlap of TF binding sites in 11 different cell lines including primary as well as cancer cells. When the centers of binding sites overlapped within 300 bp distance in all cell lines analyzed we defined them as ubiquitous; otherwise the sites were considered as cell-type specific (Methods). Among cell-type specific sites, we further classified the binding sites found in only one cell line as unique binding sites, recognizing that analysis of additional cell types might reveal these to be not truly unique. We found that overall, the MYC binding sites were highly divergent from one cell line to another, and less than 10% of MYC sites were ubiquitous, suggesting that a large portion of MYC sites in a given cell line is cell-type specific (Fig. 3-7B). Interestingly, the majority of cell-type specific binding sites of MYC were detected in ES cells. In contrast to MYC, more than half of CTCF binding sites were ubiquitous across the 11 cell lines. More than three quarters of CTCF binding sites were identified in at least 7 cell types, with less than 3 % of CTCF sites unique in any of the cell types we analyzed, except ES (6.4%) and MCF7 (3.4%) (Fig 3-7B). Similarly, Pol II also exhibited a strong binding preference for ubiquitous sites; however, unlike CTCF, a significant proportion of Pol II binding sites were also unique to a single cell

type (an average of 7.7 % across cell types). Taken together, these results suggest that MYC binding predominantly regulates unique cell type functions whereas CTCF's regulatory role is largely consistent across diverse cell types. Moreover, the unique binding sites of CTCF, MYC and Pol II had lower occupancy scores across all cell types, compared with their ubiquitous binding sites (Fig. 3-8A), suggesting that cell-type specific functions might be regulated by several TFs in a combinatory manner in which TFs cooperate to bind DNA sites particularly favorable for the lower occupancy TFs.

In order to assess the number of target genes of the unique or ubiquitous sites of CTCF, MYC, and Pol II, we assigned the downstream gene of a TF-bound promoter as its target gene. Across the 10-11 analyzed cell types, we found an average number of 91, 155, and 233 unique targets and 3,321, 185, and 8,167 ubiquitous target genes, respectively, for CTCF, MYC, and Pol II (Fig 3-8B). We also found most of the ubiquitous sites of MYC and Pol II occurred in promoters, while their unique sites were found in distal as well as intronic regions (Fig 3-8C), suggesting that the unique sites of MYC and Pol II may play roles as distal regulatory elements like enhancers. In addition, we found that the cell-type specific sites of CTCF, MYC, and Pol II lacked CpG islands compared to their ubiquitous sites (Fig. 3-8D).

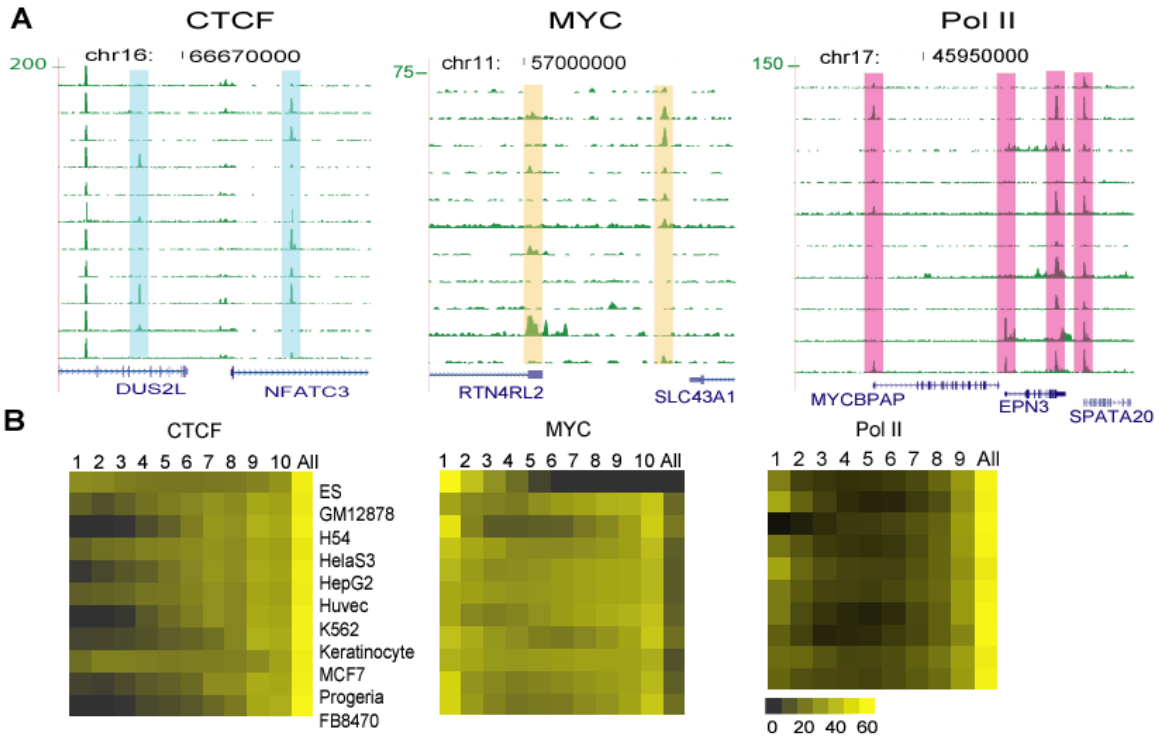

Figure 3-7. CTCF, MYC, and Pol II have many cell-type specific regulatory elements.

(A) Many cell-type specific sites were visualized in track images of genome browser. Chromosome coordinate are shown on top. Each lane shows a TF ChIP-seq track from one cell line. Gene name and location were shown in bottom track, with direction of transcription with arrows. Cell-type specific sites of CTCF, MYC, and Pol II were highlighted with color bar of sky blue, orange, and pink respectively. (B) Heat maps show the distribution of cell-type specific and ubiquitous binding sites of each factor in 10-11 different cell types. X-axis represents the number of cell types sharing a binding site across the cell types. '1' represents 'unique sites' that were found in only one cell line, 'All' indicates 'ubiquitous sites' that were found in all the cell types we examined. Cell-type specific sites are all sites except for ubiquitous sites. Y-axis represents cell types. Color bar indicates percent binding sites. The sum across X-axis is 100 % of binding sites.

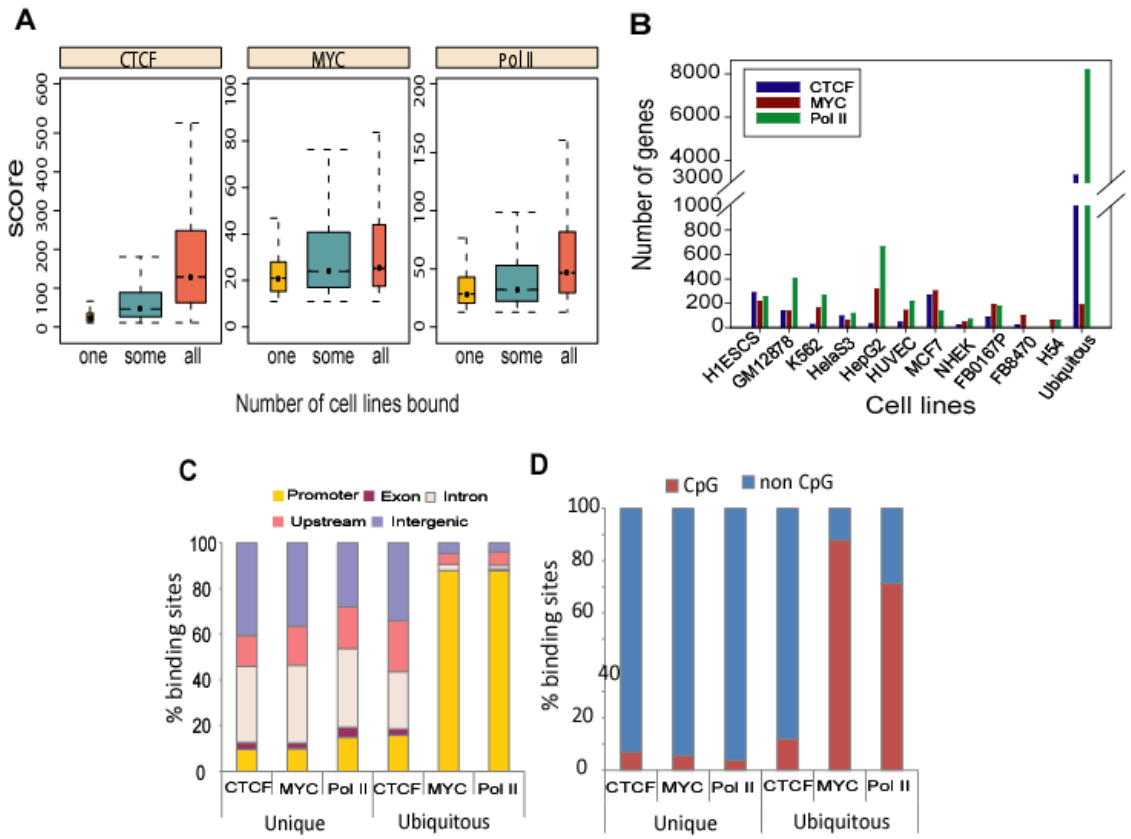

Figure 3-8. Cell type specific binding properties of CTCF, MYC, and Pol II.
(A) Boxplots exhibit the ChIP-seq score distribution of unique and ubiquitous sites across the cell types analyzed. Unique sites have significantly lower ChIP-seq score than the other sites. One, some, and all indicate unique sites, cell-type specific sites except unique ones, and ubiquitous sites, respectively. *P*-values were calculated by Wilcoxon rank sum test. Three stars (***) indicate *P*-value of zero. (B) The number of unique and ubiquitous target genes of CTCF, MYC, and Pol II in diverse cells. The downstream genes of TF-bound promoters (within ± 2 kb from TSSs) are considered as target genes. (C) The distribution patterns of unique and ubiquitous binding sites of CTCF, MYC, and Pol II across the cell types analyzed in 5 different genomic regions. X-axis represents each factor in either unique or ubiquitous sites. Y-axis shows % binding sites in the genomic regions. (D) Percent CpG and non-CpG sites in unique and ubiquitous binding sites of three factors across the cell types analyzed. X-axis represents each factor in either unique or ubiquitous sites. Y-axis indicates percent binding sites of these three factors in CpG or non-CpG sites.

In order to examine biological functions targeted by the unique and ubiquitous binding sites of the three TFs we used the Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al, 2010). In particular, unique sites frequently targeted genes in functional categories that were relevant to the biological characteristics or tissue of origin of a cell type (Table 3-3). Interestingly, even though the unique sites of Myc tended to have the lowest ChIP-seq scores in our overall analysis, this class of site targeted genes in meaningful functional categories more frequently than CTCF unique sites, which had higher scores. Moreover, the unique sites of Pol II and MYC frequently targeted overlapping functional categories in several cell types including GM12878, HepG2, HUVEC, and NHEK, suggestive of combinatorial usage of these two factors in specifying cell function (Table 3-3). Ubiquitous binding sites might be expected to target housekeeping functions; for example ubiquitous MYC binding site target genes showed moderate enrichment in translational elongation (Table 3-4), consistent with previously reported functions for Myc in regulating translation and cell growth (Boon et al, 2001; van Riggelen et al, 2010) .

Table 3-3. Functional categories enriched among target genes occupied by unique sites of CTCF, MYC, and Pol II.

			Pol II			MYC			CTCF		
	GO Biological Process Term	Total Genes	Gene Hits	Fold Enrich	FDR Q-Val	Gene Hits	Fold Enrich	FDR Q-Val	Gene Hits	Fold Enrich	FDR Q-Val
FB0167P fibroblast (Progeria)	collagen fibril organization	29	11	8.6	0.0%						
	endothelial cell differentiation	21				11	3.1	0.8%			
	response to steroid hormone stimulus	259	24	2.1	3.1%						
	hormone-mediated signaling pathway	51				19	2.2	1.3%			
	developmental growth	128	16	2.8	1.3%						
	positive regulation of developmental growth	21				10	2.8	3.1%			
GM12878 lymphoblastoid cell	immune response	630	162	2.3	0.0%	70	1.6	6.2%			
	regulation of lymphocyte activation	166	48	2.6	0.0%						
	lymphocyte activation	205	62	2.7	0.0%				47	2.1	0.2%
	mononuclear cell proliferation	45							14	2.9	2.4%
H1ESC embryonic stem cell	regulation of cell fate commitment	11	6	6.9	2.3%						
	regulation of neuron differentiation	187	30	2.0	3.5%						
	neuron migration	69				32	1.7	4.2%			
	cell differentiation in spinal cord	28							16	3.0	0.1%
	metencephalon development	42							21	2.6	0.1%
	lens morphogenesis in camera-type eye	13	6	5.9	4.7%						
	camera-type eye morphogenesis	54				27	1.8	2.8%			
H54 glioblastoma	central nervous system neuron differentiation	84	11	5.5	0.1%						
	spinal cord motor neuron cell fate specification	7	3	18.0	3.6%						
	activation of protein kinase activity	130	15	4.9	0.0%						
HepG2 hepatocellular carcinoma	lipid homeostasis	65	28	2.6	0.0%	26	3.4	0.0%			
	plasma lipoprotein particle remodeling	22	14	3.8	0.1%	12	4.7	0.0%			
	regulation of fatty acid biosynthetic process	22	13	3.5	0.3%	9	3.5	1.7%			
	steroid metabolic process	216	61	1.7	0.4%	49	1.9	0.0%			
	triglyceride homeostasis	15				10	5.7	0.0%			
	triglyceride metabolic process	48	19	2.4	1.5%						
HUVEC umbilical vein endothelial cell	blood vessel development	303	75	3.1	0.0%	89	2.8	0.0%			
	angiogenesis	183	56	3.8	0.0%	62	3.3	0.0%			
	cell-substrate adhesion	102				27	2.6	0.0%	22	2.5	0.7%
	positive regulation of smooth muscle cell proliferation	33				12	4.5	0.0%	14	4.8	0.0%
	regulation of smooth muscle cell proliferation	50	16	4.0	0.0%				16	3.7	0.1%
HelaS3 cervical carcinoma	respiratory burst	15	6	5.2	3.4%						
	superoxide anion generation	15	6	5.2	3.4%						
	anti-apoptosis	213	31	1.9	3.0%						
	regulation of B cell apoptosis	9				6	7.0	0.2%			
	positive regulation of lymphocyte proliferation	65				20	3.2	0.0%			
	digestive tract morphogenesis	29							13	4.2	0.2%
	embryonic digestive tract development	19							10	5.0	0.4%
K562 chronic myeloid leukemia	response to estrogen stimulus	132				28	2.4	0.3%			
	cytokine-mediated signaling pathway	80	20	2.3	4.9%						
	myeloid cell differentiation	107	25	2.1	3.3%						
	gas transport	16	9	5.1	0.6%						
MCF7 mammary carcinoma	gland morphogenesis	85				24	2.0	4.3%	22	2.4	0.9%
	mammary gland epithelium development	40				15	2.7	2.3%	12	2.8	4.5%
	exocrine system development	40							12	2.8	4.5%
	regulation of Rho protein signal transduction	114				35	2.2	0.1%			
NHEK epidermal cell	hemidesmosome assembly	11	7	36.9	0.0%	6	7.4	0.3%			
	cell junction assembly	58	12	12.0	0.0%	13	3.0	1.1%			
	epidermis development	193	23	6.9	0.0%	29	2.0	0.9%			
	keratinocyte differentiation	66	7	6.1	4.7%						

Table 3-4. Functional categories enriched in ubiquitous binding sites of CTCF, MYC, and Pol II.

		Pol II			MYC			CTCF		
GO Biological Process Term	Total Genes	Gene Hits	Fold Enrich	FDR Q-Val	Gene Hits	Fold Enrich	FDR Q-Val	Gene Hits	Fold Enrich	FDR Q-Val
translational elongation	105	92	1.8	0.0%	9	7.7	0.1%			
ribosome biogenesis	135	113	1.7	0.0%						
nuclear export	62	55	1.8	0.0%						
protein folding	176	131	1.5	0.0%						
regulation of ubiquitin-protein ligase activity	81	68	1.7	0.0%						
cellular protein catabolic process	314				15	4.3	0.1%			
ubiquitin-dependent protein catabolic process	273				14	4.6	0.1%			
RNA elongation from RNA polymerase II promoter	49	46	2.0	0.0%						
cellular alkene metabolic process	25							25	1.4	1.1%
unsaturated fatty acid biosynthetic process	38							35	1.3	6.4%
icosanoid biosynthetic process	35							32	1.2	10.7%
leukotriene metabolic process	24							24	1.4	1.4%
glutathione biosynthetic process	12							12	1.4	23.7%
release of cytochrome c from mitochondria	22							20	1.2	34.8%
neutral amino acid transport	20							18	1.2	45.3%

Cancer-specific sites

In order to identify cancer-specific binding sites, we also looked into cancer-specific binding sites that were present only in the cancer cell lines we investigated. We grouped the cell types into two categories: normal (GM12878, FB8470, HUVEC, NHEK and H1ESC) and cancer (HelaS3, K562, HepG2, MCF7 and H54). By comparing binding sites between these two groups, we found several cancer-specific binding sites of CTCF, MYC, and Pol II where binding was observed in all the cancer cell types but not in the normal cells. For example, Pol II occupied the promoters only in cancer cells, of 6 cancer-related genes including an isoform of PDE11A (Faucz et al, 2011; Libe et al, 2008), SATB2 (Patani et al, 2009), ALDH3B1 (Marchitti et al, 2010), SIX1 (Micalizzi et al, 2009), RAGE (Logsdon et al, 2007), and AK022914 (Zhao et al, 2007) as well as 2 more genes (PASK and MNX1) whose deregulation has not been reported in cancer (Fig. 3-9A). Interestingly, most cancer-specific binding sites were also observed in fibroblasts derived from a patient affected with Progeria, a disease characterized by rapid aging. This could reflect an underlying biological relationship between cancer and Progeria. Some of the cancer-specific binding targets showed higher expression in cancer cells compared to normal cells in accord with binding (Fig 3-9B). This raised the question of whether binding provides independent information in separating normal vs cancer, or is simply reflecting expression.

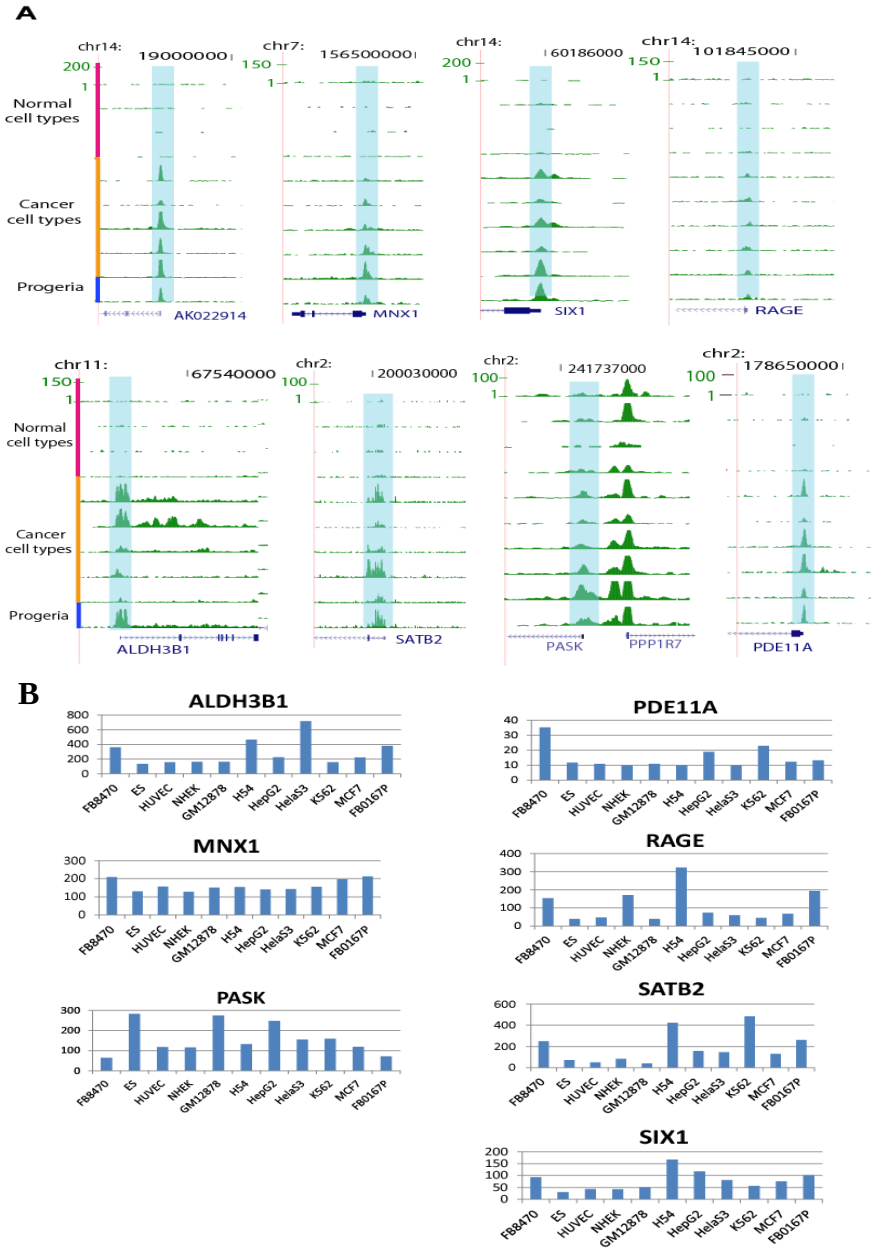

Figure 3-9. ChIP-seq revealed several cancer-specific binding sites.
(A) 8 cancer-specific sites were shown in track images of genome browser. Cancer, normal, disease (Progeria) cells were distinguished by color-coded Y-axis, pink for normal, orange for cancer, and blue for Progeria. Chromosome and coordinates are displayed on top. Cancer-specific binding sites are highlighted with sky-blue box. Gene information is shown in bottom. (B) Expression level of cancer-specific target genes in 11 different cell lines. X-axis represents cell lines and Y-axis indicates absolute expression level. Expression data were generated from Affy array.

MYC and Pol II co-localized in many promoters

To examine relationships between genes potentially targeted by TF binding, we performed clustering analysis of their target gene sets. There was generally higher correlation between the target genes of one factor across all cell types, consistent with our earlier analysis of cell-type specific and ubiquitous sites. Thus, CTCF targets generally correlated well with each other across all cell types, with Pol II targets showing lower correlations, followed by Myc (Fig. 3-10A). Between factors, there was weak, but positive correlation between CTCF and either MYC or Pol II targets (Pearson correlation coefficient r=~0.2), but moderate correlation between MYC and Pol II (r=~0.4), consistent with a functional relationship among the three factors (Fig. 3-10A).

We further investigated single or combinatorial occupancy of these factors at their target sites. We first classified binding sites into 7 groups: three single (CTCF-alone, MYC-alone, Pol II-alone) and four combinatorial (CTCF-MYC, CTCF-Pol II, MYC-Pol II, CTCF-MYC-Pol II), then examined single and combinatorial sites across cell types. Even though the largest portion (~ 86 %) of CTCF, MYC, and Pol II binding sites was bound by only one of factor, a considerable proportion of binding sites (~ 14 %) were co-localized with at least two factors in K562 (Fig. 3-10B). Similar single and combinatorial binding patterns were observed overall across other cell types. Among the combinatorial binding sites in K562, 21 % were occupied by all three factors, 10 % by CTCF-MYC, 12 %, by CTCF-Pol II and 57 % by MYC-Pol II. The fact that more than three quarters of

the co-occupied sites in this cell line were shared by MYC and Pol II reinforces the idea of a strong functional relationship between them. Associations between these factors were further supported by the fact that both CTCF and MYC were co-enriched at Pol II sites (Fig 3-10C). Although there were variations between cell types, we observed similar relationships between target genes occupied singly or in combination by these three factors (Fig. 3-11A). Taken together, these results suggest that a substantial set of genes may be regulated by combinatorial binding of these three factors, in particular Myc and Pol II.

We also examined the distribution of single or multiple-factor binding sites in the five different genomic regions relative to genes. In general, co-occupied sites were over-represented in promoters as compared to sites occupied by single factors, particularly when a combination included Pol II. 70% of the MYC-Pol II and CTCF-MYC-Pol II combinatorial sites were in promoters, which was an enrichment over the Pol2-only sites seen in promoters (Fig. 3-11B).

To globally visualize combinatorial occupancy patterns of the three factors over genes and possible functional outcomes, we first clustered Pol II occupancy signals within a 10 kb window around the TSS then visualized the corresponding signals for Myc and CTCF binding. Although dominated by the strong binding signal at the TSS, Pol II showed a few distinct clusters of binding patterns, with MYC (Fig 3-11C). These clusters were implicated in a wide range of functions. To further examine the functional relevance

of combinatorial binding, we analyzed the enrichment of functional categories among genes targeted by each of the 4 combinatorial binding groups. Target genes bound by the combination of MYC-Pol II or CTCF-MYC-Pol II showed an enrichment for genes in the functional categories of translation, RNA processing, RNA splicing, and ribosome biogenesis across all cell types, which suggests combinatorial factor usage in the regulation of genes implicated in general biological processes (Table 3-5). However, target genes occupied by CTCF-MYC or CTCF-Pol II did not show strong functional enrichment even though they exhibited moderate enrichment of some functional categories in individual cell types (Table 3-5).

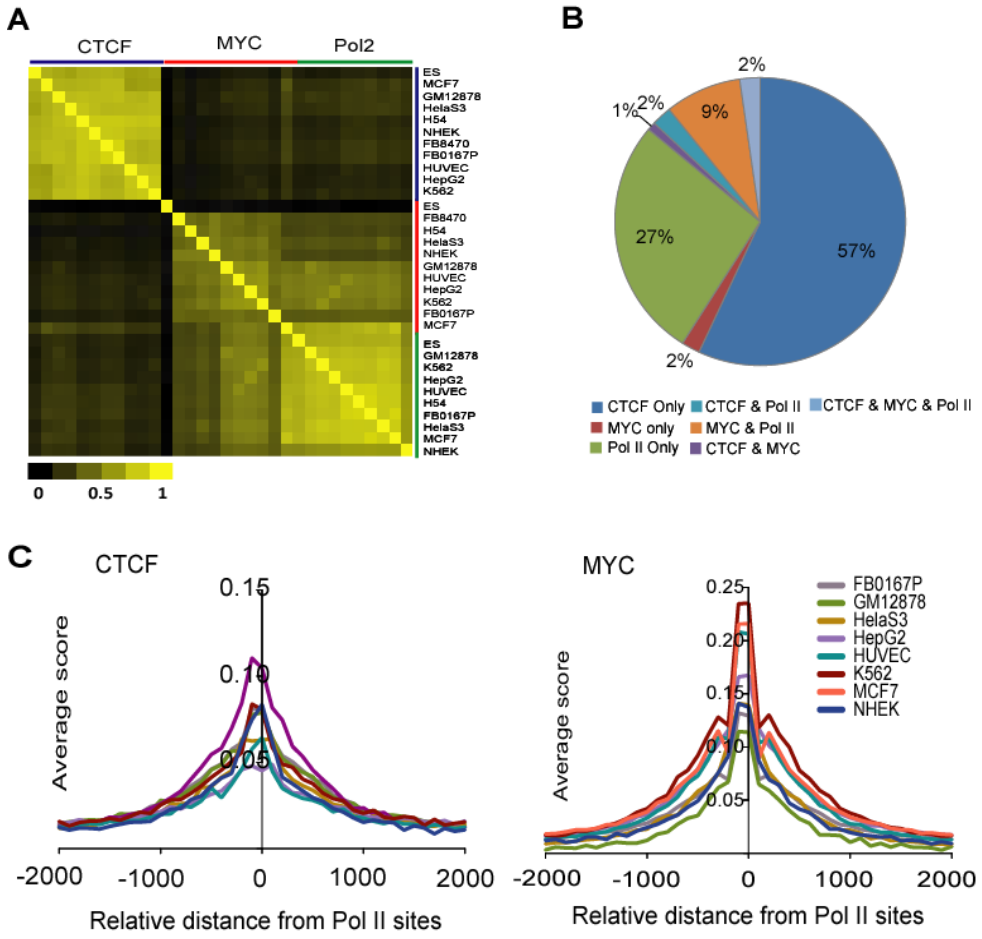

Figure 3-10. CTCF, MYC, and Pol II can regulate their target genes in a combinatorial manner. (A) A heatmap of correlations between the target genes of each factor in given cell types. The Pearson correlation coefficient was calculated between the set of target genes of each factor and those of other factors in a binary mode (target or non-target). (B) Proportion of single and combinatorial binding sites of CTCF, MYC, and Pol II in K562. (C) CTCF and MYC are co-enriched with Pol II sites

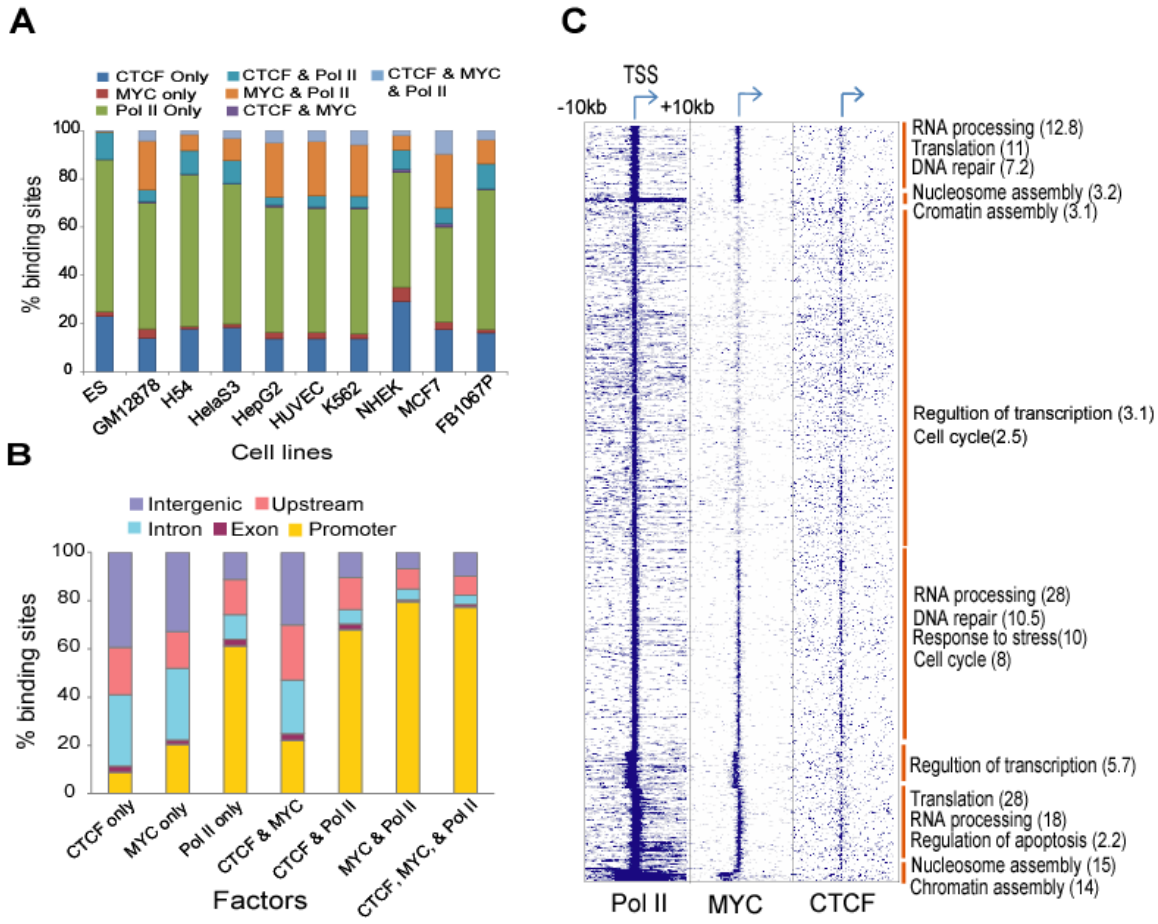

Figure 3-11. Combinatorial binding of MYC and Pol II are involved in various biological functions.

(A) Proportion of single and combinatorial target genes of the three factors in diverse cell types. The percentage of target genes in each category is shown on the vertical axis, for each of the cell types shown below. (B) The distribution of single or combinatorial binding sites of the three factors in 5 different genomic regions. The percentage of binding sites in each region is shown on the vertical axis, for each of the combinations shown below. (C) K-mean clusters show co-enrichment of CTCF and MYC with Pol II at TSS. Arrows indicate TSS as well as direction of transcription. Factors are shown in bottom. The number inside bracket indicates minus log transformed *P*-value (Bonferroni) from DAVID functional analysis. 13,300 genes are listed in Y-axis.

Table 3-5. Functional categories enriched among target genes occupied by combinations of CTCF, MYC, and Pol II.

		CTCF + MYC					CTCF + Pol II					Pol II + MYC									CTCF + MYC + Pol II								
GO Biological Process term	# gene	HelaS3	HepG2	HUVEC	K562	MCF7	GM12878	H1ESC	HepG2	MCF7	NHEK	FB0167P	GM12878	H54	HelaS3	HepG2	HUVEC	K562	MCF7	NHEK	FB0167P	GM12878	H54	HelaS3	HepG2	HUVEC	K562	MCF7	NHEK
translation	337							1.7				2.3	2.2	4.2	2.4	2.0	1.9	2.1	2.0	4.0	2.4	2.8	4.6	2.9	2.7	2.3	2.6	2.0	3.5
ribosome biogenesis	135											2.4	2.3	3.3	2.4	2.1	2.1	2.1	2.1	3.3	2.4	3.4	5.8	3.4	2.1	2.4	2.4	1.8	4.5
ncRNA metabolic process	241											2.3	2.1	3.0	2.3	2.0	1.8	2.1	2.1	2.8		2.3	2.9	2.2		1.9			
RNA splicing	301								1.8				2.0	2.6	1.8	1.8		1.8	1.9	2.6									
tRNA metabolic process	118											2.4	2.2	3.4	2.5	2.2	1.9	2.2	2.1										
tRNA processing	78											2.5			2.5		2.0	2.3	2.3	3.7									
tRNA modification	20													6.3	2.9		2.5												
tRNA aminoacylation for protein translation	45																									3.3			
protein folding	176											1.8	1.8	2.3	1.6		1.5	1.6	1.8										
'de novo' posttranslational protein folding	12					3.9																							
regulation of protein ubiquitination	123																								2.1				
positive regulation of ubiquitin-protein ligase activity	70																		1.7										
chromatin silencing	23																								4.3				
histone acetylation	53																										3.0		
histone H2A acetylation	12																									5.6			
protein amino acid acetylation	58							2.1																			2.9		
gas homeostasis	8																									7.0			
spindle organization	51																1.8												
NADP metabolic process	19																								4.6				
NADPH regeneration	12																								5.5				
glucose catabolic process	57											2.1																	
monosaccharide catabolic process	71									2.1																			
negative regulation of gene expression, epigenetic	26																										3.7		
DNA damage checkpoint	61																										3.1		
DNA integrity checkpoint	65																										2.9		
Golgi vesicle transport	138											1.8																	
post-Golgi vesicle-mediated transport	60											2.0																	
regulation of cell migration	233	2.5																											
apoptotic mitochondrial changes	30						4.0			2.7																			
regulation of transcription factor activity	144				2.7																								
regulation of binding	213	2.4			2.6																								
response to inorganic substance	270	2.3																											
actin cytoskeleton organization	257		2.1																										
icosanoid biosynthetic process	35			5.3																									
unsaturated fatty acid biosynthetic process	38			4.9																									
leukotriene biosynthetic process	20			5.9																									
cellular alkene metabolic process	25			4.7																									
positive regulation of signal transduction	292			2.2																									
SMAD protein nuclear translocation	8					5.1																							
regulation of I-kappaB kinase/NF-kappaB cascade	131						2.1																						
hemidesmosome assembly	11										6.6								6.9										
cell-substrate junction assembly	32										3.5																		

CTCF, MYC, or Pol II binding positively correlates with target gene expression

We investigated the influence of CTCF, MYC, and Pol II binding on the expression of target genes by comparing transcript levels of genes bound by any one of the factors with genes not bound by any of them. Genes whose promoters were occupied by any one of the three TFs showed significantly higher expression than genes whose promoters were not occupied by that TF, across all cell types (Fig. 3-12A). Both MYC and Pol II, compared with CTCF, showed more dramatic effects on the expression of their target genes. We could visualize the positive relationship between ChIP-seq scores and gene expression levels by plotting average binding profiles around the TSS separately for the high, medium and low expression level groups (Fig 3-12B).

In addition to the influence of TF occupancy in target gene promoters, we also investigated the effects of binding upstream (between 2 kb and 20 kb) of a TSS or within the gene-body (within exons and introns of the gene). We first assigned an upstream or gene-body binding site to the nearest gene as its target, then evaluated the expression levels of genes in each of the eight groups formed by the combination of upstream, promoter and gene-body binding by the TF (Fig 3-12C). Unlike gene body binding, upstream CTCF binding did not significantly affect target gene expression when CTCF occupied in promoter (Fig. 3-12C). Intriguingly, both upstream and gene-body binding of MYC exhibited remarkably positive effects on target gene expression only when MYC did not bind onto the promoters. Pol II upstream and gene-body binding in general were associated with higher expression regardless of its binding at the other locations. We

further looked into the influence of upstream binding distance on expression by dividing binding sites into four groups, based on distance from TSS. Interestingly, while there was a modest association of upstream binding with increased expression levels compared to no binding, there was no significant drop-off with increasing distance of binding ranging from 5 kb to 20 kb (Fig 3-12D).

Previously we showed that many promoters had a combinatorial recruitment among CTCF, MYC, and Pol II (Fig 3-12A B and D). A gene can often be transcriptionally regulated by a combinatorial binding of different transcription factors. To investigate the influence of combinatorial TF binding on target gene expression, we examined the expression level of genes whose promoters were bound by different combinations of the three factors. Genes whose promoters were bound by a single TF exhibited the highest expression levels for Pol II binding and lowest expression levels for CTCF, with MYC being intermediate (Fig. 3-12E). Genes occupied by both MYC and Pol II showed higher expression levels than the genes bound by either MYC or Pol II alone. In contrast, combinatorial binding MYC or Pol II with CTCF decreased expression level of their target genes, indicating CTCF functions as a negative regaulator of expression. These results suggest that depending on the combination of TFs, the combinatorial TF binding can either enhance or reduce the effect on expression compared to single TF binding.

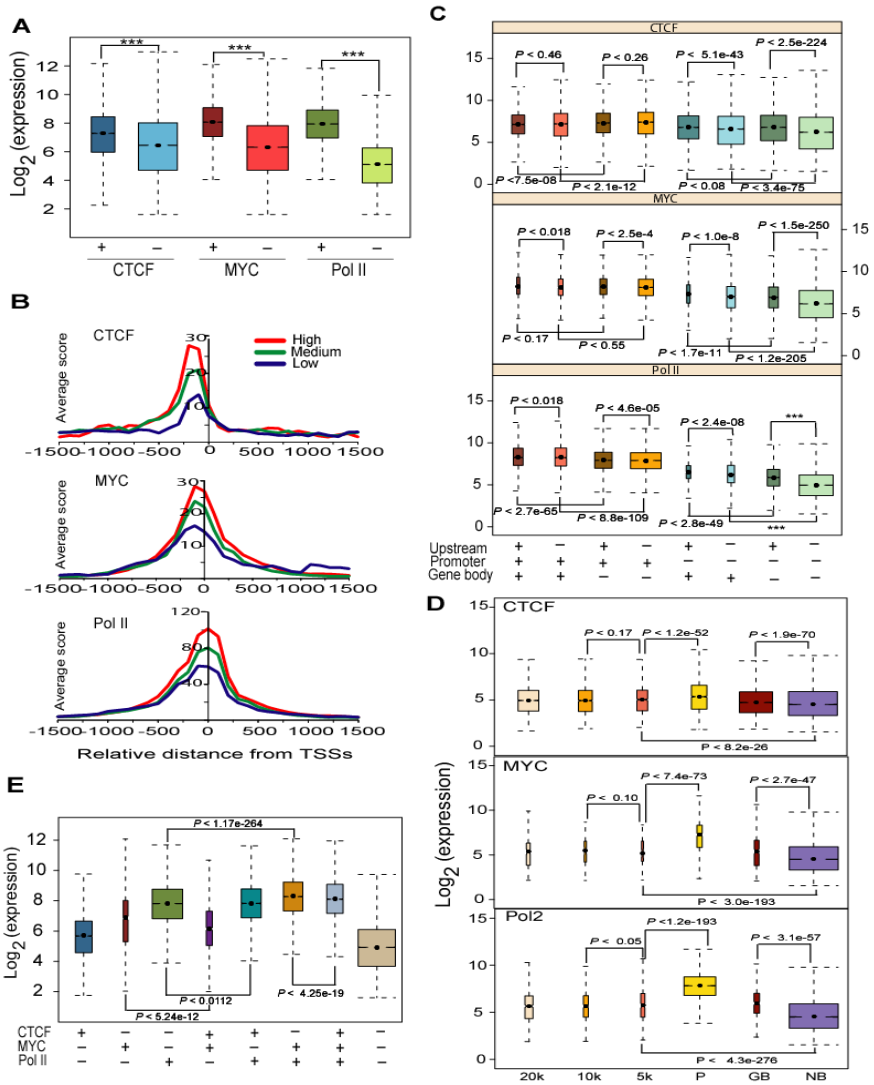

Figure 3-12. CTCF, MYC, or Pol II binding activates expression of its target genes.
(A) Boxplots show that the downstream genes of promoters (within ± 2 kb from TSSs) bound by either one of these three factors have significantly higher expression than genes not occupied by them across the cell types analyzed. *P*-values were calculated by Wilcoxon rank sum test. Three asterisks (***) indicates *P*-value of zero. (B) TSS profiles of CTCF, MYC, and Pol II among 3 different expression groups including top 33 %, middle 33 %, bottom 33 % in K562. (C) The upstream or the gene-body of a gene bound by either one of these three factors promotes its expression. Boxplots show distribution of genes expression bound by CTCF (upper panel), MYC (middle panel), and Pol II (bottom panel) in three different genomic regions including promoters, upstream, and gene bodies. *P*-values were calculated by Wilcoxon rank sum test. Three asterisks (***) indicates *P*-value of zero. (D) Location effect of CTCF, MYC, and Pol II sites on expression of target genes. *P*-values were calculated by Wilcoxon rank sum test. (E) Boxplots show expression of single and combinatorial binding of these three factors across the cell types analyzed. Combinatorial binding of MYC and Pol II enhances their target gene expression.

Pol II regulates gene expression in four distinctive binding patterns across the promoter and body of a gene

RNA polymerase II (Pol II) is an enzyme that synthesizes premature forms of mRNAs, snoRNAs, and microRNAs using DNA as a template. RNA Pol II associates not only with promoters in a simple relationship to transcription, but also shows evidence of pausing (Krumm et al, 1995). In the previous section, we showed that Pol II binding increased the expression of its target genes. Here, we investigated in further detail how location and occupancy strength of Pol II binding affect gene expression. We first classified genes into four groups based on how Pol II was bound to different gene regions: HH – showing high occupancy in both the promoter and the gene body; HL – high occupancy in the promoter and low occupancy in gene body; LH – low occupancy in promoter and high occupancy in the gene body; LL – low occupancy in both promoter and gene body (Fig. 3-13A; Methods), and then examined the expression level of each group of genes. Genes showing the HH pattern of Pol II binding had the highest level of expression whereas genes in the LL group showed the lowest level of gene expression (Fig. 3-9B). Compared to the HH group, the HL group containing paused Pol II at proximal promoters but little signal in the gene body showed significantly lower gene expression (Fig. 3-13B), which is consistent with results from Pol II profiling in *D. melanogaster* (Zeitlinger et al, 2007). To further investigate how these different modes of Pol II occupancy affected gene expression, we ranked genes by their expression values and plotted the distribution of genes in each of the four modes of Pol II occupancy as a

function of expression. The proportion of genes in the HH group decreased with decreasing expression levels while those in the LL group gradually increased (Fig. 3-13C). While genes in the HL group were distributed from highest to lowest expression, they were more biased towards highly expressed genes than genes in the LH category (Fig. 3-13C, 13D).

We next clustered genes in each of the four Pol II binding groups in order to analyze whether these distinct occupancy patterns could potentially align with functional outcomes. While there were ubiquitous as well as cell-type specific clusters in all 4 Pol II binding groups, only genes in the HH group showed strong functional enrichment across all cell types, with housekeeping functions enriched in the ubiquitous clusters where genes showed Pol II occupancy constitutively in all cell types (Fig. 3-13D). Genes in the HH group where Pol II occupancy was observed in a cell-type specific manner showed functional enrichment for cell-type specific functions, such as angiogenesis for HUVECs and lymphocyte activation in lymphoblastoid cells (Fig. 3-13D). None of the other three Pol II occupancy groups including HL, LH, and LL showed functional enrichment across all cell types, with the exception of a ubiquitous cluster in the HL group that was moderately enriched in chromatin organization/modification, DNA repair, and stress response (data not shown).

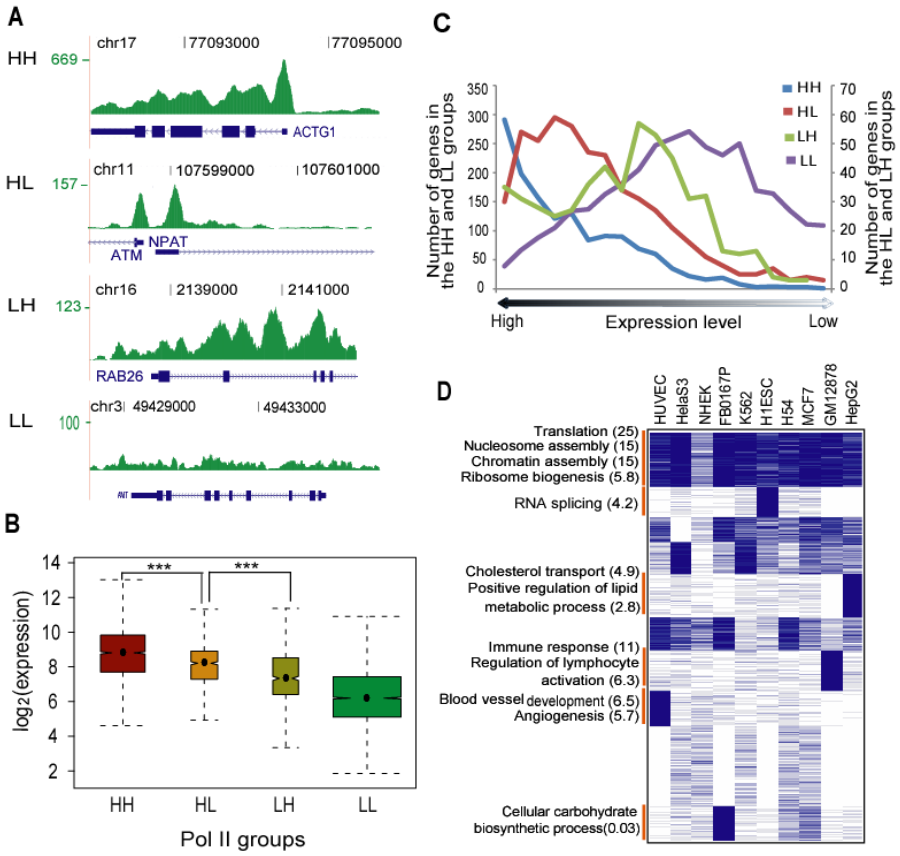

Figure 3-13. Pol II binding regulates gene expression in 4 distinct binding modes.
(A) Wiggle track images show 4 distinct Pol II binding groups classified based on its occupancy signal in the promoter and the body of a gene. HH: high occupancy in both the promoter and the body of a gene; HL: high occupancy only in the promoter of a gene; LH: high occupancy only in the body of a gene; LL: low occupancy signal in the promoter as well as the body of a gene. Chromosome and coordinates information are shown on top. Maximum occupancy signal are shown in left side of a track with a green color. (B) Boxplots shows distribution of expression level among 4 classes of Pol II binding sites. X-axis represents 4 different Pol II binding groups. Y-axis indicates log-transformed expression level of genes. *P*-values were calculated by Wilcoxon rank sum test. Three stars (***) indicates *P*-value of zero. (C) Distribution of genes of 4 Pol II groups in the expression rank map. X-axis indicates expression ranking from highest (left) to lowest (right). Y-axis represents number of genes in 4 groups, mapped into the expression map. (D) K-mean clusters shows significantly enriched functions in cell-type specific as well as ubiquitous Pol II sites in the HH group. Cell types are shown in top. Presence and absence of Pol II binding are displayed with blue and white color, respectively. Web-based functional annotation program, DAVID (Dennis et al, 2003), was used to search functional enrichment. Annotated functional groups of each cluster are shown with minus log transformed *P*-values (Bonferroni) inside bracket.

We examined in further detail the co-enrichment of Pol II with CTCF and MYC in promoters as well as gene bodies in the above 4 Pol II binding groups. CTCF showed significant enrichment only in the promoters of genes in the HH group and the gene bodies of the LH as well as LL groups while MYC exhibited highly significant enrichment in all promoters except the LL group and in the gene bodies of the HH and LH groups (Fig. 3-14), indicating that MYC occupancy at Pol II occupied genes tends to occur at gene that are strongly expressed at the level of RNA.

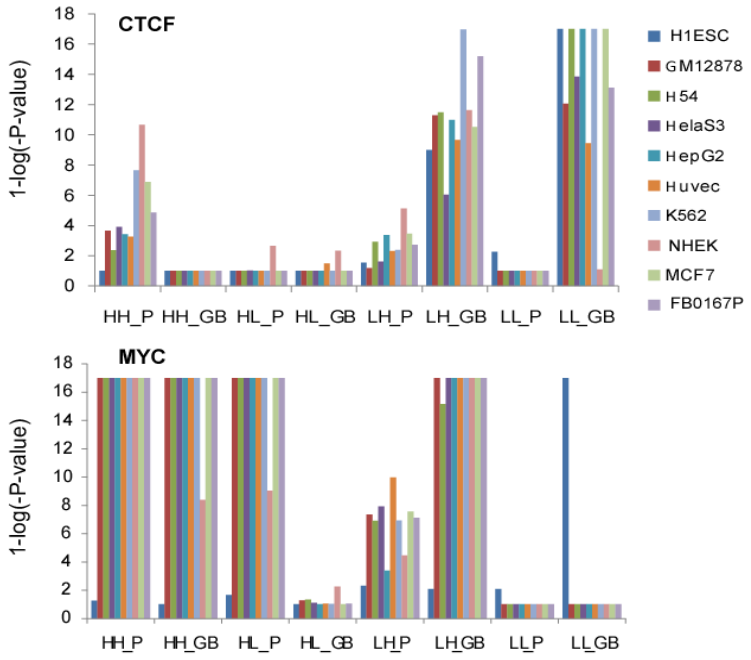

Figure 3-14. CTCF and MYC enrichment with Pol II in promoters (P) as well as gene bodies (GB) in the 4 Pol II groups.
P-values were calculated by hypergeometric distribution. X-axis represents location plus group. Y-axis shows minus log transformed *P*-values.

Novel promoters and alterative promoter usage of Pol II

From mapping of our Pol II sites into 5 different genomic regions using combined gene annotations from RefSeq, UCSC, Ensembl, Vega, and SIB genes from UCSC genome browser, we found a considerable proportion of Pol II binding sites (7~20%) in distal (upstream and intergenic) regions across the 10 cell types, suggesting the possible presence of novel promoters. In order to discriminate novel promoters from ChIP-seq false positives, we first scanned the 100 bp region around the distal Pol II binding sites for known core promoter element motifs such as initiator (INR), TATA box, TFIIB recognition element (BRE), downstream core promoter element (DPE), and motif 10 element (MTE) (Jin et al, 2006). 93 % of our distal Pol II sites contain at least 2 core promoter elements including INR, the most abundant promoter element, as well as DPE, generally found in TATA-less promoters in *Drosophila* (Fig. 3-15A). We further compared these distal Pol II binding sites with expressed sequence tag (EST) data to see whether the distal Pol II binding sites were associated with transcription. We considered a Pol II site to be associated with a transcript and thus potentially a novel promoter if it lay from 2 kb upstream to 300 bp downstream of the 5' end of the EST. On average, 74 % of the distal Pol II sites across all cell types corresponded in this manner to EST tag mRNA (Fig. 3-15B). Taken together, these results indicate that most distal Pol II sites could be promoters for novel uncharacterized transcripts rather than being truly intergenic.

Many genes in the human genome have multiple promoters for a gene corresponding to distinct transcript isoforms. These alternative promoters are differentially utilized under different cellular contexts. In order to evaluate to what extent alternative promoters are used in diverse cell types, we first identified alternative promoters for RefSeq annotated genes by flagging genes that had the same gene symbol but contained different TSS annotations in the RefFlat file. We then mapped Pol II binding sites to these alternative promoters. Fig. 3-15C shows one example of cell-type specific alternative promoter usage. Across the genome, we found that a considerable number of genes (~4.3 %) were transcribed utilizing at least two alternative promoters (Fig. 3-15D).

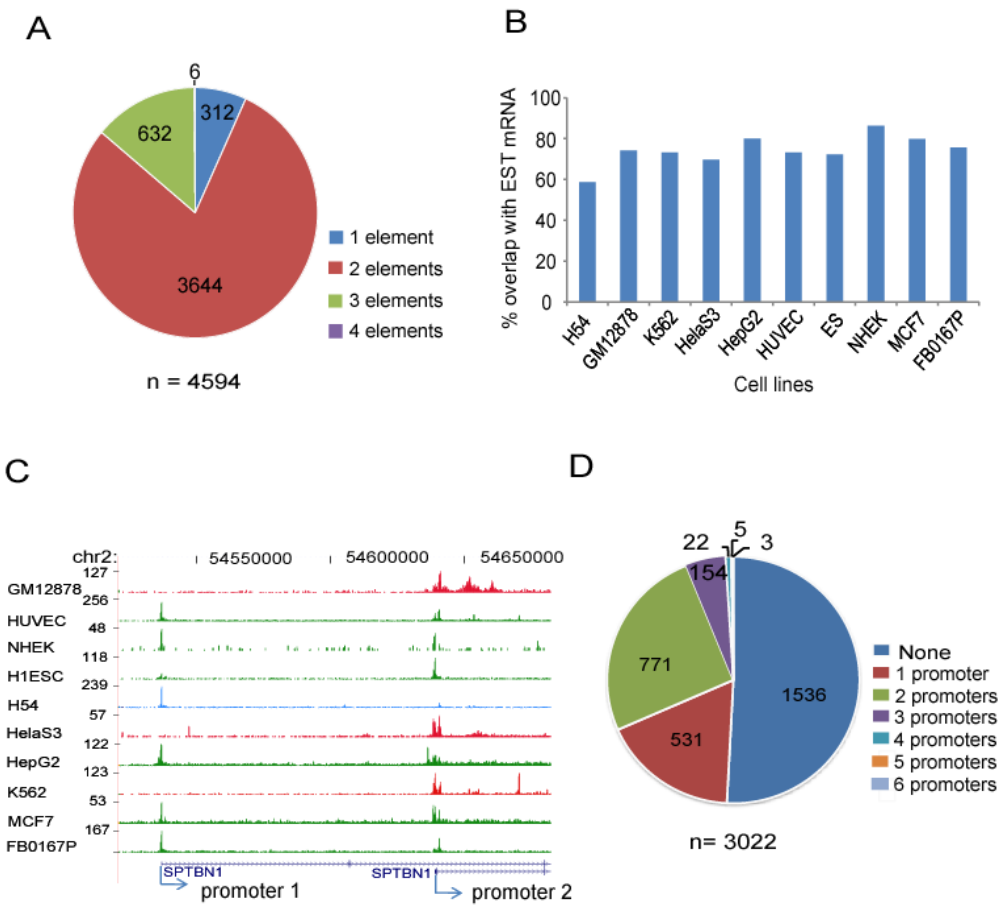

Figure 3-15. ChIP-seq of diverse cell types revealed many novel promoters as well as cell-type specific alterative promoter usage.

(A) A pie chart shows number of Pol II distal sites containing different number of core promoter elements in K562. N indicates total number of distal sites in K562, which has at least one core element. (B) Percent overlap of Pol II distal sites with expressed sequence tag (EST). (C) An example of cell type-specific alternative promoter usage is shown in genome browser tracks. Blue, red, and green indicate cell types using promoter 1, promoter 2 or both promoters respectively. Chromosome and coordinates are shown on top. Arrow indicates Two TSS locations as well as direction of transcription. Numbers in Y-axis represent a maximum peak score. (D) A pie chart represents number of genes utilizing different number of alternative promoters in K562. N indicates total number of genes having at least two promoters.

Pol II shows higher occupancy at initial and terminal exons than adjacent introns

Recent studies have noted that chromatin structure differs at exons and introns, likely reflecting an effect of co-transcriptional splicing (Schwartz & Ast, 2010). To examine whether co-transcriptional splicing might be more directly reflected in Pol II occupancy over transcripts, we examined ChIP-seq signal for Pol II at exons and introns. We first examined the Pol II enrichment around the initial and terminal exon/intron junctions. Strong Pol II occupancy around the TSS, combined with the highly variable lengths of the first exon and intron make it difficult to reliably quantify specific differences in occupancy between the first exon and intron. We visualized Pol II occupancy over the first exon/intron junction by generating gene-wise heat maps where we aligned all genes by their TSS and sorted genes by the length of their first exons. This analysis showed that in addition to the high occupancy at the TSS, Pol II also binds preferentially to the first exon compared with the intronic region downstream of it. This enrichment was seen in all 10 cell types when either input-corrected Pol II ChIP peaks (Fig. 3-16A). Similarly, to evaluate bias in Pol II occupancy at the last intron/exon junction, we aligned genes by the 5' end of their last exon and sorted them by the length of their last intron. Pol II occupancy signal was lower within the last introns while both upstream and downstream exons exhibited stronger signals (Fig. 3-16B).

To evaluate the occupancy of Pol II at internal exons, we generated heat maps of its occupancy by aligning all constitutive internal exons by their 5' ends and sorting by

the exon length. The small exon sizes and relatively sparse Pol II peaks over this region made it difficult to visualize a significant difference in occupancy between exons and intron when we considered peaks. Heat maps of raw Pol2 ChIP-seq reads revealed higher signal within the internal exons compared with adjacent regions, but a similar enrichment over the internal exons was also seen for reads from input samples (Fig. 3-16C). This enrichment is partly, but not entirely due to the greater uniqueness of sequence content within exons which results in higher alignability. To overcome the problem of sparse peaks, we combined the data for Pol2 ChIP peaks, which was corrected for input signal, to generate a combined heat map across all 10 cell types which confirmed that Pol II showed higher occupancy within internal exons (Fig. 3-16C). Thus RNA Pol II occupancy tends to be higher at exons than at adjacent introns.

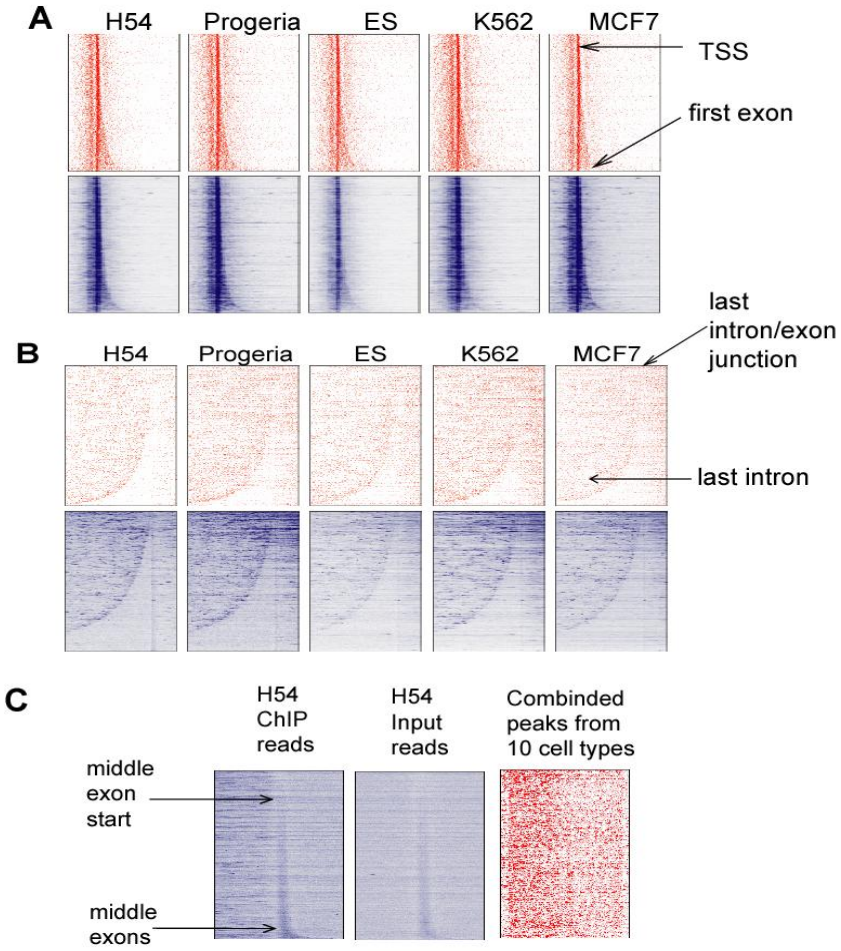

Figure 3-16. Pol II is enriched in exons.
(A) Pol II ChIP-seq peaks (top row, red) and reads (bottom row, blue) are plotted with respect to TSS. Each row indicates a gene which is aligned by its TSS and sorted by the length of its first exon. Pol2 signals in the form of input corrected peak scores (red) or read counts (blue) were assigned to 10 bp bins across the 4 kb region shown in the plots (1 kb upstream and 3 kb downstream of TSS). Only genes with at least one peak or read occurrences within 4 kb of their TSS are plotted. (B) Pol II has higher occupancy in the last exon compared with last intron. Genes are aligned with the start of their last exons and sorted by the length of their last introns. Pol2 signal across the 9kb region (7kb upstream and 2kb downstream of the start of the last exon) is represented similarly as in panel A. Most of the long genes at the bottom of the read plots (blue) do not have significant Pol2 binding peaks and therefore are not part of the peak plots (red), contributing to the pattern difference between read plots and peak plots. (C) Middle exons have higher Pol2 occupancy signal than introns. Pol2 signal across the10kb region (5kb upstream and 5kb downstream of the start of each middle exon) is represented similarly as in panel A. All these analyses were performed by Yunyun, a postdoc in Iyer lab.

Motif analysis

In order to identify sequence motifs present within the binding sites of the sequence specific transcription factors CTCF and MYC, we used the Discriminating Matrix Enumerator (DME) algorithm (Smith et al, 2005). We divided binding sites into strong, moderate, and weak groups based on their ChIP-seq scores, and searched for motifs de novo in these three groups. The algorithm identified only the previously known canonical motif for both CTCF and MYC in their respective binding sites in all cell types except in the case of MYC in ES cells (Fig. 3-17A). Next, we examined the enrichment of the motif in binding sites relative to background as a function of the binding site score for all significant binding sites of CTCF and MYC. Motif enrichment relative to background gradually increased along with ChIP-seq score for both CTCF and MYC in all 11 cell types (Fig. 3-17B).

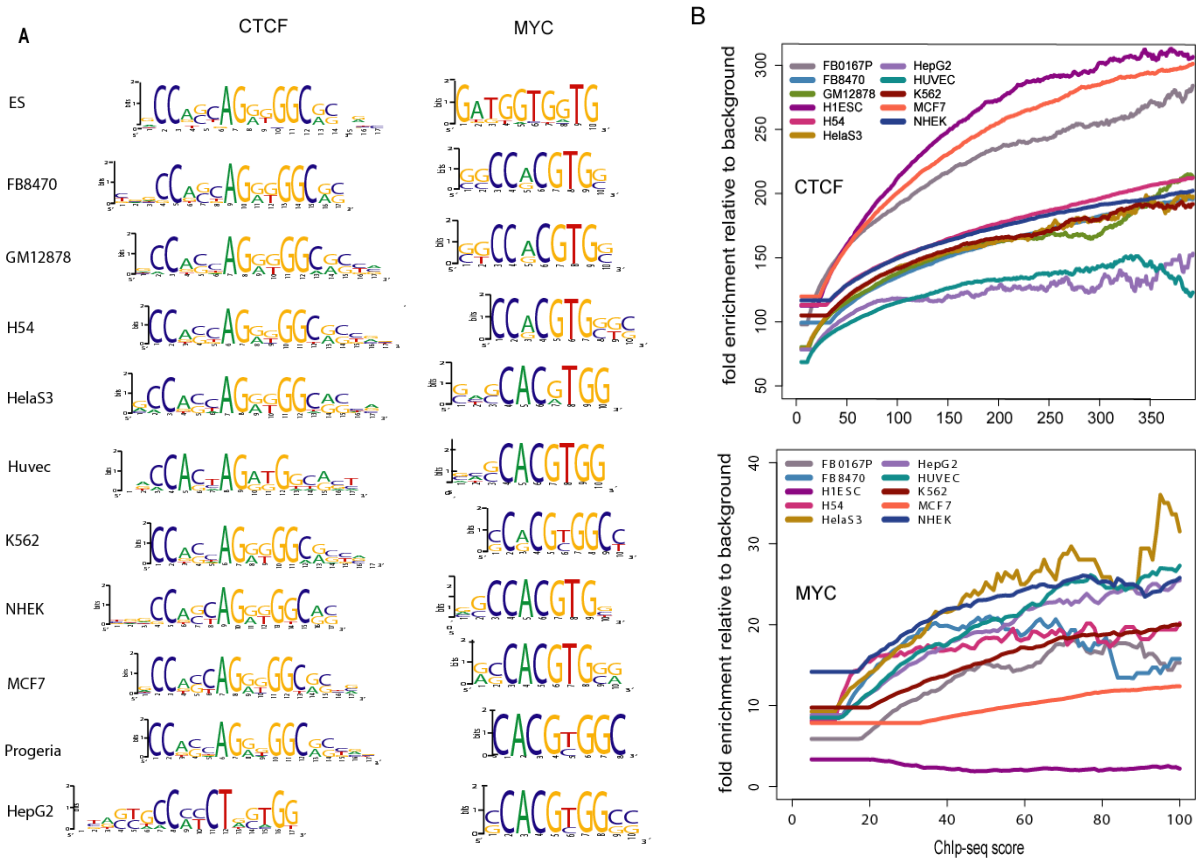

Figure 3-17. Motif analysis.
(A) Motif logo showing position weight matrix (PWM) of CTCF and MYC motifs discovered from 11 different cell types. (B) Motif enrichment relative to background along with ChIP-seq score.

3.4 DISCUSSION

We generated genome-wide high confidence binding sites of sequence-specific transcription factors including CTCF and MYC as well as Pol II in 10-11 different cell types using high-throughput ChIP-seq. The functions of TFs are governed by the location of their binding sites in genome, and the most basic regulatory mechanism is to bind onto the promoter of their target genes. In accordance with this fundamental mechanism, our genome-wide binding analysis of CTCF, MYC and Pol II showed that all three factors were overrepresented in TSS regardless of cell types. However, unlike other two factors, CTCF exhibited a distinct binding distribution across the genome where it had the largest portion of binding sites in distal sites including upstream and intergenic. These preferential CTCF binding in distal sites is consistent with its function as either an enhancer blocker which in general, locates between an enhancer and a promoter and prevents communication between them or an insulator which positions at chromatin boundaries and prevent the spreading of repressive chromatin modification signals (Cuddapah et al, 2009) . Other than the binding properties of CTCF, MYC, and Pol II we also showed that the number of binding sites of each factor positively correlated with gene density, indicates TF play functions near genes rather than gene desert. Bidirectional promoters are able to regulate expression of two downstream genes and are responsible for expression of ~11 % genes in human (Trinklein et al, 2004). It has been reported that motifs of some transcription factors including MYC E2F1, E2F4, YY1, and NRF-1 are

overrepresented in bidirectional promoters (Lin et al, 2007). Our genome-wide in vivo binding data also revealed that MYC is highly enriched in bidirectional promoters.

Through overlap analysis of binding sites of CTCF, MYC, and Pol II among 10-11 different cell types we revealed that binding sites of CTCF were highly conserved across the investigated cell types whereas those of MYC were more distinctive in each cell type, suggests global roles of CTCF and cell-type specific roles of MYC. Unlike dominant CTCF binding in ubiquitous sites, the binding sites of Pol II had not only high portion of ubiquitous sites but also considerable amount of unique sites, which implies Pol II is involved in both general and cell-type specific roles. Of cell-type specific sites and ubiquitous sites we found that ChIP-seq scores of ubiquitous sites were significantly higher than those of cell-type specific sites, in particular unique ones, implying cell type specific sites might be regulated by combination of several TFs, which cooperate with to bind onto DNA. Functional category analysis for targets of these three factors further unveiled that unique sites of MYC and Pol II had strong enrichment in representative functions of a specific cell type even though they had lower occupancy scores, but those of CTCF did not. In addition to cell-type-wise overlap analysis, we also examined factor-wise overlap, through which we revealed moderate target correlation between MYC and Pol II as well as weak but constant correlation between CTCF and MYC across cell lines analyzed, implying a functional link among these factors.

By comparing the binding sites of CTCF, MYC and Pol II with Affy exon expression array data we showed that TF binding, in general, activated expression of its target genes for all three factors and in particular, combinatorial binding between MYC and Pol II enhanced their target gene expression. These results suggest that single TF binding promotes activation of target genes and combinatory TFs binding can enhance the effect on expression. Thus, it is reasonable to speculate that the effect of TF combinations is greater than the effect of stronger binding of a single TF. Recent research revealed that MYC had a crucial role in releasing paused Pol II, (Rahl et al, 2010), which could explain why sites co-occupied with MYC and Pol II had higher expression. Moreover, we showed positive influence of CTCF, MYC, and Pol II binding in the upstream on their target genes, suggesting their binding on enhancers.

By regulating the occupancy of Pol II in different genomic regions, cells can regulate transcriptional gene expression. Our genome-wide systematic analysis of Pol II binding sites in a gene elucidated distinctive Pol II binding patterns, showing the influence of Pol II binding location on its target gene expression. In particular, although the HL group having promoter-paused Pol II showed significantly lower expression of its target genes compared to those of the HH group, genes in the HL group were still highly expressed. This result suggests that many actively transcribed genes also possess paused Pol II at their proximal promoters, but majority of Pol II pausing is transient rather than long-lasting, which keep genes from being expressed. Our observation is also in accordance with most recent genome-scale Pol II profiling studies in *D. melanogaster*

which revealed that Pol II proximal pausing is prevalent even in actively transcribed genes across the fly genome, in particular genes occupied by NELF as well as DSIF, negative effectors of transcriptional elongation (Gilchrist et al, 2010). We also found many novel promoters as well as cell-type specific alterative promoter usage, which allow cells to regulate gene expression in more diverse ways.

It has been well established that splicing takes place while pre-mRNA is still being actively transcribed (Schwartz & Ast, 2010). Without intact Pol II, splicing of pre-mRNA occurs less efficiently, implying Pol II has a role in splicing. Pol II elongation rate can be slowed down due to nucleosomes, which can function as a speed bumps (Hodges et al, 2009). Since exons have ~1.5 fold more nucleosome than introns it is reasonable to speculate that Pol II elongation rate could decrease in exons. Decreased elongation rate could increase recognition of exons and facilitates splicing (de la Mata et al, 2003; Ip et al, 2011; Schwartz et al, 2009). Previous Pol2 ChIP-ChIP studies in plants have shown that Pol2 binds to exons stronger than introns (Chodavarapu et al, 2010). However, it is not clear whether Pol2 is more enriched in exons than introns in human genome. We have shown that Pol2 enrichment is higher within exons in vivo through ChIP-sequencing in 10 different human cell lines. We compared the level of Pol2 binding difference between exons and their adjacent introns across 5 cell lines and found that the extent of exon-bias is consistent for most exon/intron pairs across different cell types. It will be interesting to correlate the expression levels and splicing status with the extent of exon-bias of Pol2

binding which should shed light on whether the Pol2 exon bias is related to co-transcriptional splicing.

Chapter 4: Allele-specific and individual-specific CTCF recruitment in the human genome

4.1 INTRODUCTION

Human has two alleles for every gene, and the regulation of a gene expression can differ between the two. Moreover, variations in genomic DNA, including single nucleotide polymorphisms (SNPs), insertion, deletion, and copy number variation (CNV), can alter gene expression in individuals (Fanciulli et al, 2007; Zeggini et al, 2007). This allele-specific gene expression can be observed in development and in many diseases (Feagins et al, 2006; Walsh & Bestor, 1999). Recent analyses have revealed that allele specificity in individuals is responsible for the expression of 10%-22% of genes in human (Zhang et al, 2009), and upwards of 30% of genes have variations in gene expression at least in part due to genetic effects (Stranger et al, 2007).

Gene expression is regulated transcriptionally by the interaction of TFs with cis-regulatory elements in the genome; however, it is not well understood how and to what extent individual differences and genetic variations affect TF binding in cis-regulatory elements and subsequent effects on gene expression. Elucidating these allele-specific and individual-specific variances in gene expression is pivotal not only to understanding the molecular basis of phenotypic variation between different individuals but also to

unveiling the causative mechanisms of many diseases that are associated with common genetic variants occurring in non-coding cis-regulatory elements.

Genome-wide identification of TF binding sites through chromatin immuno-precipitation followed by (ChIP-seq) makes it possible to investigate the extent of variation in TF binding between either alleles or individuals. Here, we provide an inclusive list of individual-specific and allele-specific variation in CCCTC binding factor (CTCF) in different human individuals. CTCF has multiple regulatory functions including activating as well as repressing genes, blocking enhancers and insulating chromatin (Zlatanova & Caiafa, 2009).

We found the presence of both allele-specific and individual-specific CTCF binding in parent-child trios from European and African populations. We also found that these allele-specific and individual-specific CTCF binding sites were inheritable from parent to child.

4.2 MATERIALS AND METHODS

Cell Line and Growth

Lymphoblastoid cell lines from the CEU (CEPH - Utah residents with ancestry from northern and western Europe) and YRI (Yoruba in Ibadan, Nigeria) sample populations were purchased from Coriell and were cultivated according to standard growth procedures (http://ccr.coriell.org). Genotype information about those individual cell lines was obtained from the HapMap and 1000 Genomes Project (20981092). Cells were grown in the RPMI 1640 (2 mM L-glutamine) medium supplemented with 15% fetal bovine serum plus 1% pen/strep. Cells were split every 3 day into fresh media. Two biological replicates were grown on separate days for each cell line.

ChIP sequencing

ChIP for CTCF was conducted using a previously described method (Kim et al, 2008). Fixed cells with formaldehyde (final concentration of 1%) were sheared with a Bioruptor into an average DNA size of 500 bp fragments. Sheared chromatin was used to pull down CTCF-DNA complex using an anti-CTCF antibody (07-729) from Millipore. After reverse crosslinking through overnight incubation in a 65 ^{o}C water bath followed by proteinase K treatment, purified ChIP DNA was used to generate ChIP-seq libraries according to Illumina's recommended protocols. Purified ChIP-seq libraries were sequenced using the Illumina GA2 Sequencer at Duke University.

Mapping and identifying peaks

Reads were mapped to the reference human genome using Maq aligner (Li et al, 2008). The reference genome excluded chromosome Y for female individuals and pseudo-autosomal regions of chromosome Y for male individuals. F-seq (Boyle et al, 2008) was used at a low threshold (4 standard deviations above the mean) to generate putative CTCF binding sites across the genome.

Gene expression analysis

RNA was extracted from lymphoblastoid cell lines at the same time they were harvested for CTCF ChIP experiments. RNA was extracted using Trizol (Invitrogen), labelled, and hybridized onto Affymetrix 1.0 exon arrays. Array CEL files were normalized with RMA. Gene level analysis was conducted using Expression Console from Affymetrix. Expression values from replicates were then averaged. To assess the binding site relevance to expression, each CTCF binding site was assigned to its nearest gene. Correlation analysis was performed with R.

Allele-specific site discovery

Alignment of reads to a single reference genome caused an artificial bias towards the reference genome in measuring allele specific binding bias in CTCF sites. To remove this bias we first reconstructed the single reference genome into two different hg18 genomes (genome1 and genome2) for each individual, based on the SNP calls obtained

from the April 2009 1000 Genomes data release. The two bases in each heterozygous SNP were randomly assigned to either genome1 or genome2 at the respective position in hg18. All reads from CTCF ChIP-seq were aligned to both genome1 and genome2 references for each individual using Maq. Approximately 5% of reads changed position in comparison to the other genomic alignment. The new mapping approach combining genome1 and genome2 improves reference genome mapping bias (Fig 4.1). The allele-specific bias for 9,192 heterozygous CTCF binding sites was assessed across all individuals. P-values were calculated based on the binomial distribution coupled with a false discovery rate (FDR) multiple testing corrections at a threshold of 0.01. 11% of CTCF sites showed significant allelic bias. All analyses to identify allele-specific and individual-specific sites were done by Ryan.

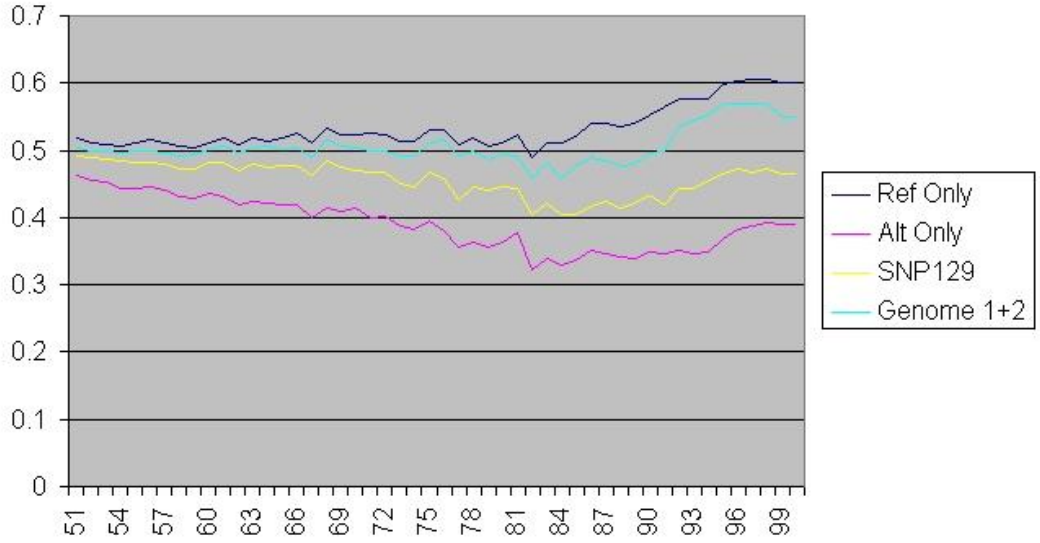

Figure 4-1. New mapping strategy removed bias toward reference allele.
X-axis represents % alignment. Y-axis represents a mapping bias. 0.5 indicates no bias. Analysis was done by Ryan.

3.3 RESULTS AND DISCUSSION

Genome-wide identification of CTCF binding sites

We generated CTCF ChIP-seq data from two independent growths for each of the 6 lymphoblastoid cells from CEU (CEPH Utah reference family) and YRI (Yoruba from Ibadan, Nigeria) family cell lines. All the samples were sequenced using Illumina Solexa sequencer through which we altogether generated over 200 million sequences for this analysis (Table 4-1). We first aligned raw sequences to the human reference genome using Maq (Li et al, 2008) followed by determination of CTCF binding sites using the F-seq package. Correlation studies using Pearson correlation showed strong agreement between each replicate of CTCF, showing that the predominant genome-wide signal was consistent with a specific biological state associated with lymphoblastoid cell lines (data not shown).

Table 4-1. Sequencing statistics of CTCF ChIP-seq

		CTCF ChIP-seq		
Family Structure	**Cell line**	**# of total sequences**	**# of useable sequences**	**% of total**
CEU Father	GM12891	30,244,488	21,733,635	71.8%
CEU Mother	GM12892	44,885,150	34,494,412	77.0%
CEU Daughter	GM12878	32,547,270	25,846,561	79.4%
YRI Father	GM19239	26,628,402	20,232,825	76.1%
YRI Mother	GM19238	32,377,472	25,547,799	78.9%
YRI Daughter	GM19240	33,399,839	26,250,278	78.4%
	Total	200,082,621	154,105,510	

Individual-specific CTCF sites are correlated between parent and child

We investigated overlap of CTCF binding sites in only the four parents that were unrelated to each other to identify individual-specific CTCF sites. We first integrated CTCF binding sites in individuals, and classified them as either "constant", meaning that they were observed in all four parent lines, or "individual-specific", meaning they were present in at least two individuals and absent in at least one individual. We found 58,192 constant CTCF binding sites and 823 CEU-specific as well as 809 YRI-specific CTCF sites (Table 4-2). We also found concordance between parent CTCF sites and child CTCF sites in each population since strong occupancy signals in the parents also tended to be strong in their child.

Table 4-2. Number of constant as well as individual-specific CTCF binding sites.

Table	Constant	CEU-specific	YRI-specific	Other combination	Singletone
# of CTCF sites	**58,192**	**823**	**809**	**7,343**	**4,117**

Influence of individual-specific CTCF binding on gene expression

We examined the functional relevance of the individual-specific sites we identified by comparing individual variations in CTCF binding sites with variations in gene expression as measured by Affymetrix exon arrays. In general, CTCF sites near the transcription start site (TSS) showed strong positive correlation with expression levels, even though there were numerous examples of sites where there was an anti-correlation (Fig. 4-2A). To explore these relationships globally, we assigned CTCF site to its nearest TSS, and then calculated the correlation between CTCF signal and the corresponding gene expression level. Many CTCF sites have both positive and negative correlated values with gene expression for sites ranging 2.5 Kb to 10 Kb of the TSS (Fig 4-2B). In addition, constant as well as individual-specific CTCF sites also showed both positive and negative correlation between CTCF binding and expression, although individual-specific site showed more positive correlation (Fig 4-2C, and D), suggesting that CTCF regulates gene expression in a more complex manner, instead of functioning just a repressor or an activator, with different roles for different genes.

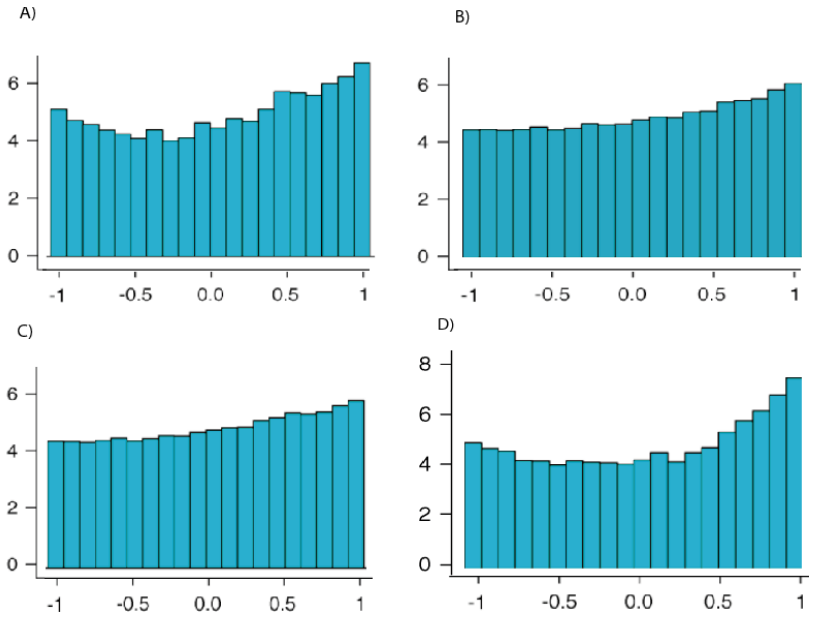

Figure 4-2. CTCF binding sites correlate with gene expression.
X-axis represents Pearson correlation coefficient and Y-axis shows percent of total. Distribution of correlation values between CTCF and gene expression across individuals. CTCF binding sites were classified into those within 2.5 Kb of the nearest TSS (A), between 2.5 Kb and 10kb from the nearest TSS (B), constant (C), or individual-specific (D). This analysis was done by Ewan Birney (McDaniell et al, 2010).

De novo identification of allelic bias on CTCF binding sites

In order to identify allele-specific CTCF binding sites where the signals from the two alleles differed significantly without reference SNP bias due to mapping toward single reference genome, we aligned sequences from our CTCF ChIP-seq to each of two reference genomes, each containing one of the possible heterozygous SNP alleles based on genotype calls for these individuals obtained from the 1000 Genomes Project (Durbin et al, 2010). This new mapping strategy eliminated the background bias from alignment to a single reference genome. We then investigated the allele-specific bias for each heterozygous SNP of CTCF binding sites where at least 15 reads resided across all individuals. Among 9,192 heterozygous CTCF sites, we found 1034 (11%) CTCF sites showed significant allele bias with a *P*-value threshold of 0.01 calculated from a binomial *P*-value followed by a false discovery rate (FDR) multiple testing corrections (Table 4-3). Interestingly, we found significantly more allele-specific bias of CTCF binding on the X chromosome in females, compared with all other chromosomes. This result may be explained by X inactivation by which CTCF binding was biased towards one allele (Fig 4-4).

Table 4-3. Number of allele-specific CTCF binding sites.

This analysis was done by Ryan (McDaniell et al, 2010).

	Allele-specific in 1 individual	Not allele-specific in 1 individual	Consistent allele-specific in 2 individuals	Opposite allele-specific in 2 individuals
# of CTCF sites	1,034	8,158	98%	2%

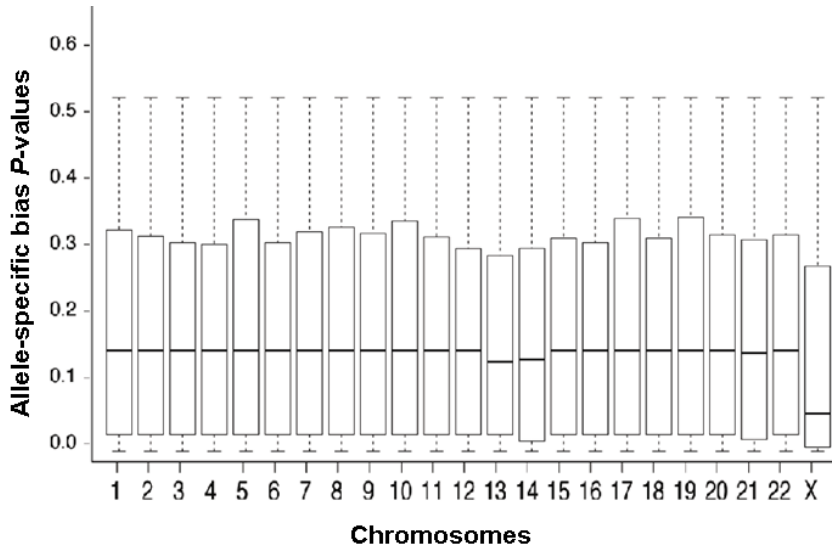

Figure 4.3. Stronger allele-specific bias on chromosome X than autosomes.
This analysis was done by Ewan Birney (McDaniell et al, 2010).

Positive correlation of allele-specificity between individuals

To test that individual-specific CTCF binding is direct consequence of genetic basis rather than random or environmental effects, including diet, infectious status at the time of isolation, and epigenetic (Hatchwell & Greally, 2007; Montgomery & Dermitzakis, 2009), we investigated the correlation of allele-specificity between individuals. We found strong correlated biases of CTCF binding to toward one of the two alleles in a heterozygous individual for the same allele between parent and child, and between individuals within and between the two populations. 98 % of inter-individual bias showed the same direction, whereas approximately less than 2% of inter-individual pairs showed statistically significant opposite behavior (Table 4.3).

Chapter 5: Summary and Future Directions

In order to identify and categorize genome-scale binding sites of sequence-specific transcription factors (TF) we performed Chromatin Immunoprecipitation (ChIP) coupled with high throughput sequencing (ChIP-seq) for E2F4, c-Myc, CTCF, and Pol II in diverse cell lines. We found tens of thousands of high confident binding sites for each factors across the human genome.

Through computational analysis using our genome-wide ChIP-seq data, we showed that all TFs analyzed had a propensity for binding near TSSs, in accordance with the fact that three TFs including E2F4, MYC and Pol II exhibited the largest portion of binding sites in promoters (among five different genomic regions: promoters, upstreams, introns, exons, and intergenics). Interestingly, unlike other TFs, CTCF showed preferential binding in distal sites including upstream and intergenic rather than promoters, which suggests a different mechanism of gene regulation (such as the well-known function of CTCF as an insulator located between promoters and enhancers, which prevents genes from activation). We also showed that the number of these TFs' binding sites positively correlated with gene density, which indicates that these factors have functions near genes rather than in gene desert.

In addition to TF binding properties, overlap analysis among 10-11 different cell lines for CTCF, MYC, and Pol II revealed many cell-type specific TFs' binding sites, which determine a cell fate. Those cell-type specific sites had lower GC contents as well

as lower binding scores. We also investigated the relationship between a TF binding and its target gene expression. In general, TF-bound genes not only had higher ChIP-seq scores but also showed higher expression than genes not bound by TFs. However, there is no clear linear correlation between ChIP-seq score and expression levels, even for Pol II binding sites.

In addition to regulating expression of target genes by associating with promoters, these TFs are able to manipulate gene expression in various other ways. For instance, our analysis of E2F4 showed that it could not only regulate microRNAs that were able to fine-tune gene expression by destabilizing translation or degrading mRNA, but also could bind distal enhancers that further activated target gene expression. We also showed in Pol II analyses that Pol II could transcribe genes using alternative promoters which could produce diverse isoforms of a given gene. Moreover, CTCF binding analysis among Europian and Yoruban family revealed allele-specific as well as individual-specific CTCF binding sites, which were inheritable from parents to their children.

To date we have generated important information about four TF binding sites across the human genome in diverse cell lines using ChIP-seq. Follow-up computational as well as experimental studies including expression arrays and luciferase assays further revealed influences of TF binding either near the promoters of genes or in distal sites. However, in many cases we found that TF binding alone was not enough to interpret changes of gene expression. In order to comprehensively understand mechanisms of gene regulation, integrated analysis is necessary: studies of other factors including

transcriptional cofactors; DNA hypersensitive site assays that search open chromatin; chromatin status investigations into histone modification as well as nucleosome position; and research into higher order chromatin structure. For instance, it is widely accepted that chromatin modification status is linked to gene expression. Genome-wide studies of histone modification have revealed many active and repressive histone marks including acetylation and methylation of a specific lysine residue of a histone, which determine open chromatin structure. One histone mark is recognized by other histone modification enzymes or histone remodelers. These sequential recruitment events result in consequential gene expression. Moreover, histone modification status can discriminate promoters enriched with H3K4me3 from enhancers with H3K4me1. Furthermore, recent genome-wide histone modification analysis has elucidated two classes of enhancers: one is an active enhancer overrepresented with H3K4me1 and H3K27Ac, and the other is a poised enhancer with H3K4me1 and H3K27me3. These poised enhancers are quiescent in ES cell, but are activated later during development by changing their histone modification status (decrease in H3K4me3 and increase in H3K4Ac). ENCODE projects have generated numerous histone modification datasets. It will be interesting to analyze ChIP seq data in conjunction with histone modification data. These combined analyses can answer many interesting questions: how many enhancer sites are regulated by a specific TF; are enhancer sites regulated by single TFs or in a combinatorial manner; and is combinatorial histone modification required for recruitment of different TFs.

Higher-order chromatin structure is another key regulatory factor in gene expression. For example, a chromatin region can communicate with another far away region by structural looping mediated by transcription factors. Chromatin conformation capture (3C) technique makes it possible to investigate such small-scale long-range integration between different genomic regions. Finally, High-C and ChIA-PET, two recently developed techniques based on high-throughput sequencing technology, provide a way to investigate genome-scale higher-order chromatin structures. It waits the next generation of methods and analyses to reveal how higher-order chromatin structure affects TFs binding of DNA, how TFs contribute to form higher-order chromatin, and to what extent these chromatin structure influences gene expression.

Appendix A. Primer sequences for qPCR as well as cloning for luciferase assay

Genes	Left	Right
EBP	ATCCCTAGTTCGGGCTCATC	CCTTCTTCGCTTCACCATTG
WDR19	CCGCGATGACTAAGATGTCA	TGCGTCTTCTTTCCTTCAGC
EID2B	GTGCCGTTATTCCAGTCTCC	GACATCTCCAACAGCCCAGT
VPS37B	GACGTCATTAATGCGCTCAC	GGGACAGTCGGGACTTCTAA
DCUN1D4	CTGGCTGGCTCTCTGCTACT	AGCTGCCTGAAAATGCACTC
CCDC41	CGCATCACTCAGACTCCAAG	GACTCAGAGATCCCCAGAGC
ZER1	ACCCGATCGCTGTTGCTAAG	CCTCCGCTGTCAACAAACC
RAD50	CAAAGCCGTAGCCACAATG	GCCTAGAGGCCCACGTGAT
WDFY2	TGGCCTAGCGGTCTTAACAA	TGCATGTTGGGAGCAGTAAG
PAPD1	ATCTTCTTTCCGGCCTCAAT	GAGGGTCAAACTAGGCGAAA
FAHD2A	TGGCAGAGAATGATTTGTGG	AGTCATCCCTCCCCTCCTC
EIF2B2	GACGGTGAACGGAAGTAACC	AGACTTGCTGCTCCCATAGC
TMEM161B	CAACTCCAGGGTGTCTGGTC	ACTCCTGCCCTCACAGAAGA
AP1B1	CCTCGCCCCACTTCTTCT	CGGGAGCTATTGGGACCT
PWP2	ACCCGGTAAGCGAACTTCAT	CCCGGGAAGTGTCTCTGTG
PHLDA3	CGTCCTAGCTTCCCAGAGC	GAACCGATCCGGAAGTGAC
TBCEL	TCACCTAGTCCCCCACTCTG	GGGTCACGTGTTGTTGTTGG
FAM55C	GCGACGTCTCCTCACCTC	GTGCTGCTGTCAGTCAACG
MTX3	GTGCCGGAAATTAGGAGGA	CCCAGCAACTGAGTTCCAA
COQ10B	GAGTCCCTCAGATGCCAAAC	GGTGCATTTCCGTGTCTTTAG
CXorf39	GACCGGAGGAGGAACTGAA	ATCCCTGATTCGATGCGTAG
CLK3	GGCCTGAGGTCTGTGTGC	AATGTGTCGTTCGCTCGTTT
ARHGAP17	TAGTAGCTGCCAGGCTGTCC	GCGGTTGAACTGCTTCTTCA
IARS	GGGATCCAGTGAAGGAGACA	GCTTGTTGGCAGGTGTCAG
AGXT2L2	GGATTTGGGGCTCAGGTTC	AAGAAATGAACCACCGCAAC
PROL1	CCAGTTTGGCAGCTTCATCT	CTTGAGCATTCTGTGTGCTG
CMTM6	GGATTCGGATGCTAAGATGC	GGGTTGACCTCAGCAGTCTC
SORT1	GGGTTGACCTCAGCAGTCTC	GGATTCGGATGCTAAGATGC
IAH1	CCACGCAGTCACTTTGGTG	GAGCTTCTTGCAAACGGACT
KLHL25	CTCTGATTGGCTGCTGCTC	ACTAGTTTCTCCGGCCTTCC
MAP2K3	GATTGGTCCTTTCGTTTCCA	TTGACAGGCAGGGACAGG
OAZ1	CCTGATTGGTGAAAGGGAAA	AGGCTCACCAACCAATCTCC
ZNF34	CTGTACACCGCTCCGTTCTC	TCTGGAGTCCGAGAAGTCAAC
RAB40B	CTGCTCAAGTTCCTGCTGGT	CAGGCTCGCCAGGATCTC

HSD17B12	AGAAGCCGCTAGTGAATGGA	TCCCAAAGTGCTGGGATTAC
SMAP2	CTTCCTCCCCACCTCCAC	CTGAAGGAAGAGCCCAGTGC
TAF1C	CTTGAGATTTCCCTCGTGGT	GTCGTGGTCGTCTGGAAAAT
ATF2	CCTTAAGCCTGACGGAATCA	CCTTCTTGCCTTCCTTCTGA
UBE2H	GTAAGCAGCCCCCTCTCAGT	GGGGGCTCAGTCACTCAC
EYA2	ATGACCCCTGTGAAAGGAAC	GGGTGGTTGAGTGAACGAAT
TUBB3	AAGAGGGGCCATTGTCCT	GAAAGGAGGGGCTGTCTCC
ETV7	GAGCGCTCAAGACAGAAAGC	GCCAGGCTCTTACCTGCAT
mir-17	GTGGGGCTTGTCCGTATTTA	AAGGACCATGTGGGTGAATG
mir-22	AGTCCTTAAAGGGCGACAGC	CGAGTCAGTTTGGGGAATGT
let-7a	AGTGAGGGGACGGACGAG	AAGCCGTCTGATTGAAGTGC
Negative control	CCGGAAGCACTTCTCCTAGA	AAGAGAGAGCGGAAGTGACG
RBL2	TTGACTCCCAGAAGGGTGAC	ATGCCTCCTTCCAAGTCCTC
TFDP-1	CGGAGAACTCAAGGTCTTCAT	GACGGTGGAGGGGTGAAC
E2F4	TGCAGAAGTCCAGGGAATG	TGAGCTCACCACTGTCCTTG
GAPDH	CTGGGCTACACTGAGCACCAG	CCAGCGTCAAAGGTGGAC

Primers for cloning distal sites to perform luciferase assay

	Left	Right
E1	GAGGTGCCCTTTGACTAGA	TACAGACAGACCCTGCCACA
E2	GGGAAGCTACCAAGGTCCAG	TAGCAAAGGGCAGAGAGAGG
E3	GGGCTCCTCTCCTAAAATGG	GGGCTCACCAGAATGGTAAA
E4	AGGAACTGGTGATGGAGGTG	GTTCAGGTGTCCGCAACTTT
E5	CCACACACACGCTCTCTGTT	AGCAGCTCAGCTCTCCTACG
E6	GCCATGGAGGTTTCTGACAT	ACCAGCTCCTGGATTCATTG
E7	AGGGGAAAACACAGCAAGTG	GAACAAAAGGACCAGGAAACC
E8	AGGGAGGGGGAAAAGAAAGT	TACGCAAACGATTCTCATGG
E9	TCGCATAATTGGGTCTTCAA	CCCCAGGTCTCTGTTGAAAT
E10	GAAAAAGTCCGGGTGACGTA	TTTGCTACCATTTGCCCAAG
Positive control	GCTCCAAGGTTAAATGTAATAGGG	AAATGTTGCAGAGAGACTCAAGTG

Appendix B. Bidirectional E2F4 binding sites

Chr	Gene	Strand	Start	Symbol	Strand	Sstart	Peak position
chr1	B3GALT6	+	1157491	SDF4	-	1157310	1157382
chr1	RER1	+	2313073	MORN1	-	2312853	2312980
chr1	TAS1R1	+	6538020	NOL9	-	6537168	6537244
chr1	C1orf213	+	23568050	ZNF436	-	23567466	23568016
chr1	ZCCHC17	+	31542428	WDR57	-	31542231	31542398
chr1	RBBP4	+	32889410	ZBTB8OS	-	32888772	32889291
chr1	CDCA8	+	37930745	C1orf109	-	37928779	37928935
chr1	C1orf50	+	43005502	LEPRE1	-	43005342	43005353
chr1	PRPF38A	+	52642806	ORC1L	-	52642719	52642735
chr1	LRRC42	+	54184624	HSPB11	-	54183876	54183900
chr1	C1orf83	+	54291861	TMEM59	-	54291699	54291718
chr1	EFCAB7	+	63761855	ITGB3BP	-	63761423	63761516
chr1	SFRS11	+	70443952	LRRC40	-	70443863	70443902
chr1	DCLRE1B	+	114249560	AP4B1	-	114249264	114249461
chr1	PRUNE	+	149247596	FAM63A	-	149245957	149246677
chr1	CKS1B	+	153213837	SHC1	-	153213583	153213781
chr1	TMEM79	+	154519362	SMG5	-	154519244	154519288
chr1	C1orf66	+	154964901	ISG20L2	-	154964329	154964885
chr1	IQWD1	+	166172531	BRP44	-	166171857	166172051
chr1	C1orf112	+	168031173	C1orf156	-	168030655	168030803
chr1	RGL1	+	181871830	ARPC5	-	181871608	181871616
chr1	KIAA0133	+	227828603	TAF5L	-	227828417	227828518
chr10	FBXO18	+	5972219	ANKRD16	-	5971352	5971923
chr10	CISD1	+	59698900	IPMK	-	59697700	59697920
chr10	UBTD1	+	99248757	MMS19	-	99248356	99248578
chr10	FBXL15	+	104169560	PSD	-	104168891	104169460
chr10	SFXN2	+	104464287	ARL3	-	104464180	104464260
chr11	RIC8A	+	198529	BET1L	-	197422	197692
chr11	LRRC56	+	527526	HRAS	-	525550	525826
chr11	RASSF7	+	551449	C11orf35	-	550779	550997
chr11	ILK	+	6581539	KIAA0409	-	6581387	6581434
chr11	C11orf17	+	8889276	ST5	-	8889074	8889264
chr11	GTF2H1	+	18300718	HPS5	-	18300297	18300492

chr11	DNAJC24	+	31347952	DCDC1	-	31347897	31347901
chr11	C11orf79	+	60954172	FLJ12529	-	60953849	60953875
chr11	TTC9C	+	62252527	HNRNPUL2	-	62251397	62251565
chr11	NUDT22	+	63750313	TRPT1	-	63750257	63750295
chr11	GPR137	+	63809906	BAD	-	63808740	63809397
chr11	CCS	+	66117265	CCDC87	-	66117130	66117240
chr11	C11orf82	+	82290384	PRCP	-	82289205	82289242
chr11	JMJD2D	+	94346492	CWC15	-	94346424	94346467
chr11	CEP57	+	95163289	FAM76B	-	95162602	95162748
chr11	JRKL	+	95763461	CCDC82	-	95762731	95762942
chr11	ATM	+	107598768	NPAT	-	107598575	107598626
chr11	RBM7	+	113776593	C11orf71	-	113776349	113776466
chr11	RNF214	+	116608613	PCSK7	-	116608021	116608063
chr11	FOXRED1	+	125644264	SRPR	-	125643960	125644197
chr11	ACAD8	+	133628643	THYN1	-	133628470	133628632
chr12	C12orf32	+	2856649	FOXM1	-	2856564	2856628
chr12	RAD51AP1	+	4518316	C12orf4	-	4517898	4518262
chr12	NCAPD2	+	6473558	MRPL51	-	6472732	6473366
chr12	USP5	+	6831545	CDCA3	-	6830717	6830746
chr12	C12orf60	+	14847772	WBP11	-	14847668	14847715
chr12	GOLT1B	+	21546066	RECQL	-	21545796	21545880
chr12	IRAK4	+	42439019	PUS7L	-	42438863	42438974
chr12	TARBP2	+	52180971	MAP3K12	-	52179538	52179883
chr12	HNRNPA1	+	52960754	CBX5	-	52960182	52960185
chr12	C12orf26	+	81276454	CCDC59	-	81276330	81276388
chr12	VEZT	+	94135652	FGD6	-	94135371	94135405
chr12	APAF1	+	97563208	IKIP	-	97562720	97562934
chr12	SCYL2	+	99185679	DEPDC4	-	99184988	99185506
chr12	C12orf48	+	101038085	NUP37	-	101036491	101038054
chr12	ISCU	+	107480423	SART3	-	107479295	107480334
chr12	MVK	+	108495882	MMAB	-	108495741	108495759
chr12	ERP29	+	110935534	TMEM116	-	110935318	110935390
chr12	C12orf52	+	112107937	DDX54	-	112107667	112107675
chr12	RNFT2	+	115660478	C12orf49	-	115660226	115660451
chr12	DYNLL1	+	119392042	SFRS9	-	119391941	119391987
chr12	KNTC1	+	121577761	RSRC2	-	121577500	121577721
chr12	ZNF664	+	123023622	CCDC92	-	123023116	123023343

chr12	NOC4L	+	131194945	DDX51	-	131194833	131194944
chr12	PXMP2	+	131774264	POLE	-	131774018	131774047
chr13	EXOSC8	+	36472916	ALG5	-	36471477	36471502
chr13	SETDB2	+	48916510	CAB39L	-	48916222	48916297
chr13	ALG11	+	51484550	ATP7B	-	51483631	51484111
chr13	TMCO3	+	113193308	DCUN1D2	-	113193024	113193042
chr14	RNF31	+	23686498	PSME2	-	23685695	23686365
chr14	C14orf21	+	23838937	DHRS1	-	23838506	23838833
chr14	LTB4R	+	23852356	CIDEB	-	23850416	23850490
chr14	FANCM	+	44674885	FKBP3	-	44674272	44674854
chr14	ATP5S	+	49848796	L2HGDH	-	49848697	49848724
chr14	GPR137C	+	52089615	TXNDC16	-	52088963	52089282
chr14	SOCS4	+	54563593	WDHD1	-	54563557	54563564
chr14	SLC38A6	+	60517632	TRMT5	-	60517535	60517622
chr14	TTLL5	+	75197373	C14orf1	-	75196912	75197320
chr14	C14orf174	+	76913514	TMED8	-	76913149	76913277
chr14	KIAA1409	+	92869317	BTBD7	-	92869138	92869255
chr14	FAM14B	+	93617391	DDX24	-	93617311	93617376
chr14	CCNK	+	99017491	SETD3	-	99016979	99017106
chr14	BTBD6	+	104786052	BRF1	-	104785267	104785886
chr15	CHP	+	39310728	EXDL1	-	39310187	39310704
chr15	DTWD1	+	47700587	C15orf33	-	47700410	47700437
chr15	IQCH	+	65334241	FLJ11506	-	65334128	65334133
chr15	LRRC49	+	68972040	THAP10	-	68971807	68971932
chr16	C16orf33	+	43828	POLR3K	-	43625	43717
chr16	NUBP2	+	1772933	SPSB3	-	1772582	1772582
chr16	UBN1	+	4837912	N-PAC	-	4837304	4837414
chr16	ASPHD1	+	29819647	SEZ6L2	-	29818081	29818945
chr16	CCDC95	+	29915031	HIRIP3	-	29914888	29914929
chr16	RNF40	+	30681130	LOC90835	-	30681043	30681096
chr16	ORC6L	+	45281058	VPS35	-	45280645	45281046
chr16	PHKB	+	46052710	ITFG1	-	46052516	46052524
chr16	FBXL8	+	65751391	TRADD	-	65751313	65751313
chr16	SLC9A5	+	65840355	FHOD1	-	65838926	65839074
chr16	SF3B3	+	69115201	COG4	-	69114958	69115198
chr16	ZC3H18	+	87164344	C16orf85	-	87164049	87164212
chr16	C16orf55	+	88251710	CHMP1A	-	88251630	88251667
chr17	RPA1	+	1680022	SMYD4	-	1679925	1680017

chr17	TMEM93	+	3518838	TAX1BP3	-	3518722	3518820
chr17	RNF167	+	4784374	SLC25A11	-	4784063	4784100
chr17	MIS12	+	5330970	DERL2	-	5330218	5330631
chr17	POLR2A	+	7328573	ZBTB4	-	7328292	7328343
chr17	WDR79	+	7532519	TP53	-	7531588	7532423
chr17	CNTROB	+	7776197	TRAPPC1	-	7775983	7776122
chr17	PFAS	+	8093361	C17orf68	-	8092138	8093295
chr17	C17orf39	+	17883335	ATPAF2	-	17883205	17883295
chr17	SMCR8	+	18159318	TOP3A	-	18159046	18159049
chr17	PIGW	+	31965515	MYO19	-	31964838	31964895
chr17	TUBG1	+	38015219	FAM134C	-	38014928	38015192
chr17	CNTD1	+	38204379	CCDC56	-	38204230	38204274
chr17	TMUB2	+	39619879	C17orf65	-	39619608	39619674
chr17	CCDC103	+	40332679	EFTUD2	-	40332289	40332390
chr17	UTP18	+	46692895	MBTD1	-	46692426	46692878
chr17	PRR11	+	54587874	FAM33A	-	54587582	54587826
chr17	CCDC45	+	59933619	DDX5	-	59932869	59933529
chr17	TMEM104	+	70284216	NAT9	-	70284065	70284117
chr17	TSEN54	+	71024203	CASKIN2	-	71023222	71023687
chr17	SAP30BP	+	71174993	RECQL5	-	71174864	71174990
chr17	MFSD11	+	72245377	SFRS2	-	72245007	72245158
chr17	TMC8	+	73638453	TMC6	-	73636456	73637397
chr17	CCDC137	+	77244165	C17orf90	-	77244023	77244106
chr17	ANAPC11	+	77442894	THOC4	-	77442758	77442879
chr17	LRRC45	+	77574568	STRA13	-	77574062	77574353
chr18	NDC80	+	2561604	METTL4	-	2561489	2561538
chr18	PSMG2	+	12693063	CEP76	-	12692703	12692724
chr18	RNMT	+	13716703	C18orf19	-	13716591	13716653
chr18	KIAA1468	+	58005503	PIGN	-	58005269	58005361
chr18	C18orf55	+	69966725	FBXO15	-	69965929	69965979
chr18	TSHZ1	+	71051718	ZADH2	-	71050105	71050536
chr19	REEP6	+	1442164	PCSK4	-	1441407	1441427
chr19	SF3A2	+	2187815	PLEKHJ1	-	2187328	2187773
chr19	SAFB	+	5574163	SAFB2	-	5573938	5574083
chr19	TMEM146	+	5671687	LONP1	-	5671176	5671224
chr19	BTBD14B	+	13090108	TRMT1	-	13088332	13089996
chr19	CC2D1A	+	13878051	C19orf57	-	13877909	13877932
chr19	TMEM38A	+	16632937	C19orf42	-	16631968	16632133

chr19	MYO9B	+	17047590	NY-SAR-48	-	17047343	17047565
chr19	RFXANK	+	19164007	MEF2B	-	19163933	19163977
chr19	C19orf40	+	38154987	CCDC123	-	38154709	38154909
chr19	PSENEN	+	40928333	U2AF1L4	-	40928176	40928259
chr19	C19orf55	+	40940883	HSPB6	-	40939770	40940825
chr19	WDR62	+	41237622	THAP8	-	41237504	41237609
chr19	EIF3K	+	43801561	MAP4K1	-	43800483	43801518
chr19	SNRPA	+	45948618	C19orf54	-	45947668	45948533
chr19	BCKDHA	+	46595543	EXOSC5	-	46595096	46595103
chr19	ZNF576	+	48792383	IRGQ	-	48791516	48792283
chr19	BLOC1S3	+	50373842	TRAPPC6A	-	50373325	50373398
chr19	QPCTL	+	50887771	SNRPD2	-	50887282	50887344
chr19	SYNGR4	+	53559468	TMEM143	-	53558998	53559396
chr19	RUVBL2	+	54188967	GYS1	-	54188361	54188432
chr19	ZNF524	+	60803541	FIZ1	-	60802705	60803352
chr2	TTC15	+	3362452	TSSC1	-	3360605	3360691
chr2	CENPO	+	24869836	C2orf79	-	24869755	24869818
chr2	C2orf13	+	68548245	FBXO48	-	68547894	68548005
chr2	CCT7	+	73314912	C2orf7	-	73313864	73313893
chr2	DOK1	+	74635367	LOXL3	-	74634570	74634829
chr2	RBED1	+	85435446	RETSAT	-	85435166	85435414
chr2	CIAO1	+	96295610	TMEM127	-	96295459	96295528
chr2	EIF5B	+	99320265	TXNDC9	-	99319292	99320181
chr2	DBI	+	119840973	C2orf76	-	119840728	119840861
chr2	ARL6IP6	+	153283375	PRPF40A	-	153282221	153282290
chr2	PKP4	+	159021721	CCDC148	-	159021460	159021643
chr2	NIF3L1	+	201462390	PPIL3	-	201462244	201462362
chr2	EEF1B2	+	206732562	NDUFS1	-	206732432	206732542
chr2	RQCD1	+	219141921	USP37	-	219141328	219141594
chr2	FAM134A	+	219751182	C2orf24	-	219749946	219749993
chr2	EIF4E2	+	233123600	TIGD1	-	233123470	233123502
chr2	PPP1R7	+	241738574	PASK	-	241737551	241737621
chr20	CDS2	+	5055481	PCNA	-	5055268	5055269
chr20	MCM8	+	5879297	TRMT6	-	5879173	5879288
chr20	C20orf72	+	17897761	SNX5	-	17897490	17897647
chr20	POFUT1	+	30259356	PLAGL2	-	30259207	30259345
chr20	RPRD1B	+	36095361	KIAA0406	-	36095247	36095318
chr20	CSTF1	+	54400833	AURKA	-	54400758	54400770

chr20	C20orf177	+	57948950	PPP1R3D	-	57948747	57948894
chr20	C20orf11	+	61039885	DIDO1	-	61039719	61039752
chr21	SON	+	33837219	GART	-	33836286	33836318
chr21	SH3BGR	+	39739666	LCA5L	-	39737998	39739625
chr21	RRP1B	+	43903859	HSF2BP	-	43903802	43903852
chr21	C21orf57	+	46530694	MCM3AP	-	46529664	46530554
chr22	MRPL40	+	17800035	HIRA	-	17799219	17799346
chr22	CDC45L	+	17847415	UFD1L	-	17846726	17846755
chr22	RANBP1	+	18485023	HTF9C	-	18484768	18484900
chr22	SNAP29	+	19543291	PI4KA	-	19543070	19543268
chr22	HSCB	+	27468042	CHEK2	-	27467822	27467837
chr22	MPST	+	35745647	TST	-	35745437	35745640
chr22	XPNPEP3	+	39583039	ST13	-	39582633	39582990
chr22	CRELD2	+	48698347	ALG12	-	48698110	48698220
chr3	PARP3	+	51951400	RRP9	-	51950962	51951248
chr3	IL17RB	+	53855616	CHDH	-	53855216	53855460
chr3	ATXN7	+	63825272	THOC7	-	63824637	63825187
chr3	SLC35A5	+	113763584	ATG3	-	113763175	113763289
chr3	FAM162A	+	123585712	CCDC58	-	123584764	123585686
chr3	DIRC2	+	123996590	HSPBAP1	-	123995340	123995362
chr3	IFT122	+	130641657	MBD4	-	130641542	130641609
chr3	SMC4	+	161600123	IFT80	-	161600014	161600034
chr3	POLR2H	+	185563887	CLCN2	-	185562085	185562300
chr3	FBXO45	+	197780121	WDR53	-	197779810	197779936
chr3	PIGX	+	197923642	C3orf34	-	197923520	197923580
chr3	LMLN	+	199171467	IQCG	-	199171271	199171313
chr4	TMEM175	+	916261	GAK	-	916174	916195
chr4	C4orf42	+	1234176	CTBP1	-	1232908	1233136
chr4	TACC3	+	1693063	TMEM129	-	1692882	1693025
chr4	GRK4	+	2935140	NOL14	-	2934916	2935093
chr4	MGC21874	+	7096056	CCDC96	-	7095629	7095660
chr4	NCAPG	+	17421622	C4orf30	-	17421479	17421591
chr4	ENOPH1	+	83570749	HNRPDL	-	83570402	83570588
chr4	CISD2	+	104009575	UBE2D3	-	104009473	104009502
chr4	LARP7	+	113777568	C4orf21	-	113777505	113777541
chr4	SPATA5	+	124063674	NUDT6	-	124063573	124063662
chr4	ARFIP1	+	153920561	TIGD4	-	153920327	153920359
chr5	SDHA	+	271355	CCDC127	-	271297	271304

chr5	EXOC3	+	496333	LOC116349	-	496210	496252
chr5	TRIP13	+	946003	BRD9	-	945915	945987
chr5	NDUFS6	+	1854508	MRPL36	-	1852956	1853047
chr5	SRD5A1	+	6686499	NSUN2	-	6686157	6686341
chr5	C5orf22	+	31568129	RNASEN	-	31568039	31568075
chr5	SKIV2L2	+	54639332	DHX29	-	54639278	54639323
chr5	PPWD1	+	64894886	CENPK	-	64894751	64894777
chr5	PTCD2	+	71651955	MRPS27	-	71651840	71651860
chr5	XRCC4	+	82409072	TMEM167A	-	82408935	82409070
chr5	ANKRD32	+	93980146	C5orf36	-	93980040	93980095
chr5	RELL2	+	140996700	HDAC3	-	140996607	140996678
chr5	HMMR	+	162820240	NUDCD2	-	162819721	162820233
chr6	THEM2	+	24775253	TTRAP	-	24775094	24775110
chr6	HIST1H2BH	+	26359857	HIST1H3F	-	26358814	26359156
chr6	HIST1H2AG	+	27208799	HIST1H2BJ	-	27208554	27208779
chr6	HIST1H2AI	+	27883955	HIST1H2BL	-	27883688	27883860
chr6	HIST1H2BM	+	27890800	HIST1H2AJ	-	27890497	27890609
chr6	TCF19	+	31234281	CCHCR1	-	31233994	31234279
chr6	PSMB9	+	32929915	TAP1	-	32929726	32929898
chr6	SYNGAP1	+	33495824	CUTA	-	33494043	33494141
chr6	NFYA	+	41148687	C6orf130	-	41148166	41148306
chr6	KLHDC3	+	43089954	MEA1	-	43089596	43089645
chr6	POLH	+	43651855	XPO5	-	43651642	43651840
chr6	MAD2L1BP	+	43705256	GTPBP2	-	43704914	43705126
chr6	CENPQ	+	49539054	MUT	-	49538990	49538992
chr6	DOPEY1	+	83834103	UBE2CBP	-	83832264	83832342
chr6	ORC3L	+	88356561	RARS2	-	88356440	88356524
chr6	C6orf182	+	109523048	SESN1	-	109521970	109522817
chr6	PEX3	+	143813809	ADAT2	-	143813534	143813658
chr6	MRPL18	+	160131481	TCP1	-	160130725	160130754
chr7	CBX3	+	26207623	HNRNPA2B1	-	26206938	26206966
chr7	CPSF4	+	98874498	PTCD1	-	98874355	98874436
chr7	AP4M1	+	99537065	MCM7	-	99536316	99536438
chr7	CNPY4	+	99555200	TAF6	-	99554915	99555149
chr7	MEPCE	+	99865464	ZCWPW1	-	99864238	99864345
chr7	LRWD1	+	101892394	ALKBH4	-	101892293	101892338
chr7	DNAJB9	+	107997591	THAP5	-	107997403	107997538
chr7	SLC4A2	+	150387589	CDK5	-	150385929	150386108

chr8	ESCO2	+	27687976	CCDC25	-	27686089	27686099
chr8	ADAM9	+	38973661	TM2D2	-	38973198	38973236
chr8	MCM4	+	49036046	PRKDC	-	49035296	49035995
chr8	CHCHD7	+	57286868	PLAG1	-	57286413	57286781
chr8	PTDSS1	+	97343342	MTERFD1	-	97342972	97343011
chr8	POP1	+	99199243	HRSP12	-	99198594	99198711
chr8	ENY2	+	110415811	NUDCD1	-	110415526	110415728
chr8	KIFC2	+	145662545	CYHR1	-	145661839	145662500
chr8	LRRC14	+	145714198	RECQL4	-	145714008	145714060
chr9	IFT74	+	26937309	PLAA	-	26937139	26937238
chr9	NUDT2	+	34319503	KIF24	-	34319198	34319283
chr9	DNAI1	+	34448810	C9orf25	-	34448568	34448731
chr9	CREB3	+	35722316	TLN1	-	35722128	35722277
chr9	OSTF1	+	76893217	C9orf95	-	76892953	76893093
chr9	SLC31A1	+	115023688	FKBP15	-	115023462	115023588
chr9	C9orf43	+	115212842	POLE3	-	115212773	115212816
chr9	MRRF	+	124066967	RBM18	-	124066911	124066933
chr9	SLC25A25	+	129870299	NAIF1	-	129869212	129870271
chr9	LRRC8A	+	130684211	CCBL1	-	130684175	130684185
chr9	C9orf163	+	138497767	SEC16A	-	138497328	138497349
chr9	C8G	+	138959518	FBXW5	-	138958994	138959000
chr9	NDOR1	+	139220003	TMEM203	-	139219911	139219972
chrX	MOSPD2	+	14801483	FANCB	-	14801105	14801133
chrX	KIF4A	+	69426619	PDZD11	-	69426523	69426591
chrX	TMEM187	+	152891184	HCFC1	-	152890013	152890109
chrX	IKBKG	+	153428755	G6PD	-	153428427	153428713

Appendix C. TF targets of E2F4 from ChIP-seq

TF	ChIP score	TF	ChIP score	TF	ChIP score
ZFPL1	159.09	ZNF764	8.19	LITAF	5.74
TRIP13	129.83	TRIM27	8.16	NR2F2	5.74
BRD9	129.83	TRIM27	8.16	OPTN	5.72
FOXM1	99.85	KLF6	8.15	ZNF354A	5.72
TCF19	96.65	ZNF684	8.15	PCGF2	5.72
LRRC14	92.92	RNF4	8.13	OSR2	5.72
CTCF	90.35	ZBTB44	8.11	KLF9	5.68
HMX2	89.87	C14orf166	8.09	IRF2	5.68
UHRF1	73.63	PER2	8.09	RNF141	5.68
ZNF653	69.64	TAF6	8.08	NUFIP2	5.68
ZNF688	66.69	CTNNB1	8.07	ZNF420	5.68
C14orf106	62.05	EGR1	8.07	MSX1	5.66
HMGB2	54.51	ZNF256	8.07	NAB1	5.66
TIMELESS	47.06	ULK2	8.05	ZFAND5	5.66
C15orf42	46.82	ELK1	8.04	GRHL1	5.66
RBL1	46.58	MYST2	8.03	BCL11A	5.64
DMTF1	46.05	ZNF84	7.99	ZNF582	5.63
MYBL2	45.14	ZNF687	7.99	SNAPC3	5.62
E2F3	44.55	RFXANK	7.96	ZNF202	5.62
SLC25A40	44.03	MEF2B	7.96	DLX2	5.62
ASH2L	42.69	TULP3	7.95	CNOT3	5.61
SUV39H1	40.1	ZNF138	7.95	JMJD1C	5.6
IRF8	39.89	NFAT5	7.94	PHF17	5.6
JMJD2D	39.69	HIC1	7.93	NR2F6	5.59
E2F2	37.06	PAWR	7.92	ZNF691	5.59
HDAC4	37.02	ENO1	7.91	RUNX3	5.58
ZNF331	35.02	YY1	7.91	BACH1	5.58
MTF2	34.34	ZNF282	7.89	ZNF323	5.57
IRF3	34.2	HCFC2	7.89	ZNF32	5.56
YEATS4	33.01	KLF13	7.89	UBP1	5.56
NFKBIL2	32.7	ZNF589	7.87	CDR2L	5.56
SART3	32.15	RNF8	7.86	SHOX2	5.55
NFYA	31.93	FOXK2	7.84	TAF1A	5.55
E2F1	31.61	SREBF2	7.82	LARP1	5.55

MYBL1	31.35	ZNF695	7.82	HSF1	5.54
ZNF689	30.42	HDAC1	7.81	ZNRF1	5.54
RFX2	29.88	ELF2	7.81	NR1H3	5.53
ARID3B	28.37	CUX1	7.75	ZNF236	5.53
FOXN2	27.75	ZNF267	7.74	ZNF214	5.53
SUPT4H1	27.13	ZNF318	7.73	PHF21A	5.53
WHSC1	25.63	BTBD6	7.73	ZNF44	5.52
MYB	24.59	ZNF646	7.72	ZNF655	5.52
TFDP1	24.47	ZNF668	7.72	MYCL1	5.51
ZNF473	24.32	IRF7	7.71	GFI1	5.51
AATF	24.14	SETBP1	7.71	ZNF623	5.51
TCF15	23.58	ATF5	7.7	GTF2E1	5.5
YAF2	23.32	ARNTL	7.64	SCMH1	5.5
ZNF519	22.21	RING1	7.63	CITED2	5.49
TAF5	22.07	SSRP1	7.62	GTF2I	5.45
HMGB3	21.28	TBPL1	7.62	AHR	5.45
SP4	21.22	ZNF827	7.62	CEBPA	5.45
HLTF	21.03	TRIM13	7.61	ECD	5.44
HMGN2	20.81	HDAC2	7.6	VSX1	5.44
FIZ1	20.78	JUN	7.59	INSM2	5.44
ZNF524	20.78	SUPT16H	7.58	EP300	5.43
EED	20.11	TRIM24	7.57	POU2AF1	5.43
YBX2	20.02	NFYC	7.57	ZNF643	5.43
ATF7IP	18.48	ZNF276	7.57	ZNF57	5.43
BLOC1S1	18.43	REL	7.56	IRF1	5.41
PIAS4	18.05	DBP	7.55	JMJD1B	5.41
EZH2	18	FOXO1	7.55	MXD1	5.4
ZFAT	17.81	HIVEP2	7.55	ZNF490	5.4
ZNF770	17.48	ZNF497	7.53	SP1	5.4
CNBP	17.26	HEY2	7.52	ATOH7	5.4
KLHL12	17.16	CREG1	7.51	ZNF791	5.4
ZNF266	17.07	RNF115	7.51	ZNF773	5.4
ZNF443	17.01	CBX6	7.48	ZNF581	5.39
DIDO1	16.48	ATG4B	7.46	ZNF768	5.39
SPEN	16.46	ZFAND3	7.46	SMAD6	5.38
SAFB	16.37	FOSL2	7.45	SOX12	5.38
ZNF766	16.36	MEF2A	7.44	PRDM4	5.38

CNPY3	16.32	ETV2	7.43	MED14	5.36
TRIM33	16.31	HPCAL1	7.43	BARHL1	5.36
IER2	15.95	ZNF398	7.42	BHLHB2	5.35
BTAF1	15.94	ZNF100	7.42	ZHX2	5.35
CIAO1	15.94	GABPA	7.41	GTF2H3	5.34
NR4A2	15.91	ZNF324	7.41	KLF7	5.34
ZNF274	15.82	ZXDC	7.4	ZNF671	5.34
RFX1	15.63	SMARCA4	7.38	ZNF3	5.33
CBX3	15.6	ZNF133	7.38	ZNF3	5.33
NR2C2	15.56	IRF5	7.32	CDK5	5.31
MYC	15.52	CAND1	7.31	HOXC8	5.31
ZNF184	15.45	NFE2L3	7.3	DLX1	5.3
ZNF180	15.42	ZNF273	7.3	ZNF703	5.29
HMGB1	15.3	ETV5	7.28	CIC	5.28
ZNF436	15.16	MESP1	7.28	CSRP2	5.27
ZNF341	14.89	BPTF	7.26	NCOA4	5.27
TFEB	14.88	NFKBIA	7.24	C19orf6	5.27
STAT1	14.79	ZNF74	7.23	ISL2	5.24
MXD3	14.74	ADNP	7.23	TP53BP2	5.23
ZBTB1	14.7	TCF12	7.22	C19orf28	5.23
ZBTB25	14.7	HMG20A	7.2	PRDM15	5.22
NR6A1	14.56	KLF16	7.2	MANSC1	5.2
EME2	14.44	IKZF3	7.19	TFB1M	5.19
RUNX1	14.38	SMARCE1	7.17	IKZF1	5.18
ZNF785	14.32	ZNF140	7.17	ZXDB	5.18
HSF2BP	14.29	ZNF174	7.15	SAP30BP	5.18
ZNF786	14.25	ETV3	7.15	SRF	5.17
POU2F1	13.96	ZNF434	7.15	RERE	5.17
CDR2	13.92	BRD7	7.1	SMARCC2	5.16
GTF3C5	13.8	NCOA2	7.08	TEF	5.16
TCERG1	13.65	ZFP36L2	7.07	NEUROG3	5.16
PBX4	13.61	SMAD2	7.05	ZNF574	5.15
ZNF567	13.49	ZEB1	7.05	LHX4	5.14
TTLL4	13.45	ZNF584	7.05	KLF5	5.13
ZNF101	13.42	ZNF10	7.04	TRIP10	5.13
GTF3A	13.35	USF2	7.03	HOXA11	5.13
ZNF200	13.35	TAF15	7.01	ZNF277	5.13
KLF10	13.34	SETD1A	7.01	ZNF526	5.13

CSRP1	13.26	TSC22D2	6.96	FOXH1	5.12
RBL2	13.19	HOXB5	6.95	E4F1	5.12
MNX1	13.1	NFKB2	6.93	INSM1	5.11
ZNF107	13.1	TTLL5	6.92	RNF114	5.11
MAZ	13.09	HOXA1	6.91	AFF1	5.1
CBX5	13.06	BANP	6.91	SNAI1	5.09
CREB3	13.04	MLL	6.9	ZFHX3	5.08
HEXIM1	13.03	JARID2	6.87	TFB2M	5.08
C16orf80	12.81	THRAP3	6.87	ZNF7	5.07
PLAGL2	12.75	NANOG	6.87	MYCBP	5.07
CBX1	12.57	GTF3C2	6.86	MLL3	5.07
SMAD4	12.44	SAP30	6.83	ESRRA	5.06
PIAS1	12.43	MORF4L2	6.83	ZNF514	5.06
TP53	12.37	ZNF337	6.82	ZFP3	5.06
PRDM2	12.36	KLHDC5	6.82	TRIM26	5.05
MXD4	12.13	ZBTB2	6.82	SOX10	5.05
ZNF76	12.03	PAX6	6.81	PRDM16	5.05
ZNF93	12.02	RBBP9	6.81	AKAP9	5.05
PBX3	11.94	ZFP1	6.79	SOX9	5.04
ADPGK	11.94	FOXN4	6.79	TADA2L	5.04
TRIM4	11.83	ZNF22	6.77	ZNF416	5.03
IRF4	11.69	FOXF1	6.76	ZNF416	5.03
NFE2L2	11.62	RUFY3	6.76	PKNOX1	5.02
ZNF672	11.56	MBD1	6.74	SUPT3H	5.01
ZNF568	11.55	FOS	6.73	MEF2D	5.01
RB1	11.53	EGR3	6.72	CEBPD	5
TRIM28	11.45	HOXC6	6.72	TOX	5
HMGA1	11.43	HOXC4	6.72	SERTAD2	4.99
ZNF692	11.42	HOXC5	6.72	CBX8	4.99
ZNF18	11.33	BCLAF1	6.69	MAFG	4.97
MBD2	11.32	TRIM9	6.69	BRPF1	4.96
BRD2	11.27	MYBBP1A	6.68	RORA	4.96
DPF2	11.2	NEUROG2	6.68	GTF2F2	4.93
MLXIP	11.18	MICALL1	6.68	SATB2	4.93
DPF1	11.12	SMAD5	6.67	ZNF146	4.92
HIRA	11.11	RBAK	6.63	MYST4	4.91
PATZ1	11.11	ETS2	6.62	ZNF322A	4.91
JUND	11.09	ZNF160	6.61	GLI4	4.91

OTP	11.07	CBFA2T2	6.61	HNF1B	4.9
PHTF1	11.06	ZNF160	6.61	SIM1	4.88
TAF5L	11.05	HOXB2	6.58	NFIL3	4.88
ZNF670	10.99	ZNF212	6.54	ZNF593	4.88
ZNF509	10.98	ZFHX4	6.54	LEF1	4.87
SIAH1	10.96	TRIM39	6.54	L3MBTL3	4.87
FOSB	10.94	CROCC	6.5	ZNF431	4.87
ZNF382	10.88	SOX4	6.48	MAX	4.86
HES1	10.75	SF1	6.48	NMI	4.86
MFSD3	10.75	ZMYND11	6.48	PHF16	4.86
ZNF696	10.72	MLL4	6.46	OTUD7B	4.86
CBX7	10.7	ZNF148	6.46	POU6F1	4.85
XBP1	10.68	PIAS3	6.45	SPIB	4.85
ARID1A	10.64	ZNF195	6.45	PDLIM5	4.85
C19orf25	10.56	SETD4	6.43	L3MBTL	4.85
ZNF294	10.45	IGHMBP2	6.42	ZNF396	4.85
REV3L	10.42	SP2	6.38	FOXJ1	4.84
MGA	10.4	ZNF33A	6.36	ZSCAN29	4.84
BRD3	10.33	JMJD2C	6.36	ZNF141	4.83
SIX5	10.33	ZNF268	6.35	ZBTB17	4.83
SNAPC4	10.3	ZNF264	6.34	KLF2	4.83
HHEX	10.28	RBBP5	6.34	KLF2	4.83
RCOR3	10.28	RASSF7	6.33	ZNF746	4.83
TP73	10.19	SMAD1	6.32	ZNF746	4.83
ASPH	10.17	TFAM	6.31	HNF4G	4.82
ZNF239	9.98	ZNF426	6.31	MXI1	4.82
ZNF430	9.97	MZF1	6.3	ACVR2A	4.81
DRAP1	9.79	ATF4	6.3	ZSCAN22	4.79
ZNF446	9.7	NFE2L1	6.28	TFAP4	4.78
PSIP1	9.7	CLOCK	6.27	ZFP36	4.78
NR4A3	9.7	ZNF639	6.27	NFYB	4.78
RNF24	9.69	LAS1L	6.27	SNW1	4.78
KEAP1	9.67	FHL2	6.26	NKX2-4	4.77
TBC1D10B	9.66	HEY1	6.26	ZMYM4	4.76
ZNF800	9.66	TSHZ1	6.26	MIER3	4.76
ZNF41	9.65	LZTR1	6.26	MTA1	4.75
ZNF143	9.63	PHF15	6.26	NR2C1	4.74
GTF2A1	9.61	ZNF347	6.26	PAX3	4.74

TAF12	9.6	POU2F3	6.25	HES6	4.73
ZNF142	9.58	ZNF445	6.24	MIXL1	4.73
NR1D2	9.57	ARID1B	6.23	KBTBD7	4.73
LMO4	9.57	TRMT1	6.22	CEBPB	4.71
ZFP62	9.56	ATXN2	6.2	DLX4	4.7
TAF4B	9.55	PER1	6.19	MYF6	4.7
MED26	9.53	MTA2	6.19	SMAD7	4.7
PMF1	9.53	CCT4	6.18	RORB	4.7
ZNF85	9.52	KLHL21	6.18	ASCL1	4.68
MLLT1	9.48	ZNF767	6.18	FOXG1	4.67
TRIP4	9.47	CHD4	6.17	ELK4	4.67
SCAND1	9.47	MED7	6.17	FLI1	4.66
LHX2	9.42	MAFF	6.17	MYOD1	4.66
SND1	9.41	ARFGAP2	6.17	ZNF613	4.65
MEIS2	9.37	MLLT10	6.16	CHD1	4.64
JUNB	9.36	CSDA	6.16	RNF144A	4.64
PLAGL1	9.35	ZNF675	6.15	ZXDA	4.63
UBR4	9.35	FOXN3	6.12	RNF13	4.63
TAF1B	9.34	ZNF706	6.12	FOXD1	4.62
ZNF384	9.33	GRHL3	6.12	ASCL2	4.62
ARIH2	9.33	ZNF576	6.12	BAZ1A	4.62
ZNF215	9.33	EP400	6.11	ZNF23	4.62
ERMP1	9.22	NR3C1	6.09	METTL3	4.61
TRIP11	9.21	ZNF544	6.09	TRIP6	4.6
TAF3	9.19	ZNF606	6.07	ELK3	4.6
ZNF287	9.18	ZNF75A	6.06	SOX13	4.6
NFX1	9.14	TAF4	6.05	POU4F1	4.6
BRF1	9.13	ZNF248	6.05	SAMD4B	4.58
SMARCA2	9.09	NFATC1	6.05	EPAS1	4.57
ZNF410	9.06	STAT5A	6.04	MNAT1	4.56
CREB3L4	9.04	HOXB7	6.04	IVNS1ABP	4.56
ARID3A	9.03	ZNF225	6.04	ZNF669	4.56
ZFP91	9.01	EGR2	6.03	POU3F1	4.54
TRIOBP	8.99	FOXD2	6.03	ATF1	4.54
INTS4	8.98	DMRT2	6	ATF1	4.54
TBX6	8.97	PURA	6	ILF3	4.54
ATOH8	8.97	TRIM14	6	IKZF2	4.54
CREB1	8.91	PRDM10	5.99	RNF135	4.53

TSC22D1	8.89	BTBD1	5.97	HSF4	4.52
PREB	8.88	MESP2	5.97	NFKB1	4.51
RFX3	8.87	ZNF92	5.95	DKFZP434B0335	4.51
ARNTL2	8.87	ZBTB4	5.95	ZFP64	4.51
MBD4	8.85	NR2F1	5.94	ZNF167	4.51
ZNF681	8.85	ZSCAN2	5.94	HAND2	4.5
DR1	8.82	LSR	5.93	AFF2	4.49
ID3	8.82	TP53I13	5.93	FOSL1	4.49
RELA	8.8	ZNF664	5.91	ZNF24	4.49
ZNF682	8.8	ZNF664	5.91	ZNF701	4.49
ZNF605	8.8	ZFX	5.9	HOXB9	4.49
ZDHHC17	8.74	DEAF1	5.89	ZNF419	4.49
ZNF263	8.73	PTTG1IP	5.87	TAF10	4.48
ZNF295	8.73	MAP3K8	5.87	JDP2	4.48
ZNF575	8.72	MORF4L1	5.87	GTF2F1	4.47
SUV420H1	8.69	MIER1	5.87	STAT4	4.47
TMEM175	8.67	RLF	5.85	AFF4	4.47
TCF3	8.66	ZBTB20	5.84	E2F5	4.46
SMARCA5	8.66	ZNF329	5.84	ARNT2	4.46
SMARCC1	8.62	ZFAND6	5.83	USF1	4.45
MNT	8.6	REPIN1	5.82	RNF10	4.45
FOXO3	8.58	MLLT3	5.81	IFT172	4.45
HMX3	8.54	NCOR1	5.81	RRN3	4.45
GABPB1	8.52	NFIC	5.81	ZNF557	4.44
JMJD2B	8.51	TBP	5.8	ZNF784	4.44
STAT5B	8.48	UBTF	5.8	SRCAP	4.43
CHD1L	8.45	PAX5	5.8	KLF11	4.42
HIF1A	8.44	ZNF26	5.8	KLHL25	4.42
LDB1	8.42	ZNF8	5.8	ZNF136	4.41
ZNF219	8.34	CIZ1	5.78	CNOT8	4.41
CEBPG	8.31	SRXN1	5.77	MED15	4.41
BHLHB5	8.31	PLAG1	5.76	PRR7	4.41
RNF2	8.29	PACS2	5.76	SCML4	4.41
SP3	8.24	TSC22D4	5.75	VDR	4.4
JMJD4	8.24	ARID5B	5.75	TEAD3	4.4
ZNF335	8.2	CRIP1	5.74	CREBBP	4.4
PIAS2	8.19	CREM	5.74	PROP1	4.4

Appendix D. Putative miRNA targets of E2F4.

Chr	Start	End	score	miRNA	miRNA position
chr5	54504758	54504850	12.7	hsa-mir-449a	chr5:54502121-54502202 (-)
chr10	105982121	105982204	14.24	hsa-mir-609	chr10:105968543-105968626 (-)
chr10	105981964	105982140	5.82	hsa-mir-609	chr10:105968543-105968626 (-)
chr10	98581759	98581823	6.62	hsa-mir-607	chr10:98578421-98578501 (-)
chr10	98582592	98582816	4.79	hsa-mir-607	chr10:98578421-98578501 (-)
chr11	61491733	61491815	8.33	hsa-mir-611	chr11:61316538-61316611 (-)
chr12	61282889	61283115	5.05	hsa-let-7i	chr12:61283728-61283825 (+)
chr12	12770213	12770260	5.64	hsa-mir-613	chr12:12808850-12808939 (+)
chr13	90797881	90798063	11.33	hsa-mir-17	chr13:90800863-90800941 (+)
chr13	49554127	49554209	14.92	hsa-mir-16-1	chr13:49521111-49521195 (-)
chr13	40261526	40261650	4.51	hsa-mir-621	chr13:40282915-40282992 (+)
chr13	98650837	98651006	11.66	hsa-mir-623	chr13:98806391-98806479 (+)
chr14	99843038	99843167	5.27	hsa-mir-345	chr14:99843956-99844033 (+)
chr15	39739987	39740139	10.4	hsa-mir-626	chr15:39771096-39771163 (+)
chr15	39739851	39740028	8.26	hsa-mir-626	chr15:39771096-39771163 (+)
chr15	53487779	53487920	7.39	hsa-mir-628	chr15:53452434-53452510 (-)
chr15	29295474	29295559	6.9	hsa-mir-211	chr15:29144544-29144621 (-)
chr15	62125569	62125671	6.11	hsa-mir-422a	chr15:61950185-61950272 (-)
chr16	15644523	15644636	5.65	hsa-mir-484	chr16:15644614-15644690 (+)
chr16	65818331	65818474	6.1	hsa-mir-328	chr16:65793721-65793803 (-)
chr17	72245107	72245234	32.28	hsa-mir-636	chr17:72244133-72244213 (-)
chr17	26910823	26910964	4.79	hsa-mir-193a	chr17:26911138-26911213 (+)
chr17	26909855	26909956	5.22	hsa-mir-193a	chr17:26911138-26911213 (+)
chr17	1901174	1901264	5.03	hsa-mir-132	chr17:1899963-1900040 (-)
chr17	7078080	7078289	4.76	hsa-mir-324	chr17:7067341-7067417 (-)
chr19	14501324	14501404	11.89	hsa-mir-639	chr19:14501355-14501447 (+)
chr19	10689668	10689852	8.33	hsa-mir-638	chr19:10690085-10690183 (+)
chr19	50834435	50834526	4.82	hsa-mir-330	chr19:50834097-50834178 (-)
chr19	57464594	57464647	16.36	hsa-mir-643	chr19:57476873-57476953 (+)
chr21	25856204	25856387	6.86	hsa-mir-155	chr21:25868156-25868236 (+)

chr3	161600209	161600344	9.19	hsa-mir-15b	chr3:161605087-161605166 (+)
chr3	161599954	161600101	7.35	hsa-mir-15b	chr3:161605087-161605166 (+)
chr5	167939108	167939216	6.08	hsa-mir-103-1	chr5:167920477-167920558 (-)
chr5	148717717	148717845	11.91	hsa-mir-143	chr5:148788690-148788764 (+)
chr6	33284091	33284289	4.56	hsa-mir-219-1	chr6:33283600-33283682 (+)
chr6	45453737	45453879	5.01	hsa-mir-586	chr6:45273404-45273480 (-)
chr6	30647074	30647143	10.84	hsa-mir-877	chr6:30660078-30660180 (+)
chr6	126702912	126702994	101.59	hsa-mir-588	chr6:126847475-126847552 (+)
chr8	22158338	22158438	5.96	hsa-mir-320	chr8:22158423-22158496 (-)
chr8	105670591	105670720	11.13	hsa-mir-548a-3	chr8:105565778-105565855 (-)
chr9	20674086	20674209	11.1	hsa-mir-491	chr9:20706109-20706184 (+)

Appendix E. Putative miRNA targets of E2F4 discovered from ChIP-seq.

Chr	peak position	score	miRNA	Strand	Distance
chr1	153431317	9.03	hsa-mir-92b	+	275
chr1	153431134	7.73	hsa-mir-92b	+	458
chr1	94084516	8.53	hsa-mir-760	+	460
chr1	94084892	5.12	hsa-mir-760	+	84
chr1	154658080	5.02	hsa-mir-9-1	-	1235
chr1	1083231	4.77	hsa-mir-200b	+	9116
chr1	1083231	4.77	hsa-mir-200a	+	9875
chr10	104182388	9.09	hsa-mir-146b	+	3871
chr10	112730754	5.04	hsa-mir-548e	+	7920
chr10	21829592	4.81	hsa-mir-1915	-	4016
chr11	61341051	12.09	hsa-mir-1908	-	1763
chr11	566462	9.64	hsa-mir-210	-	8264
chr11	61339371	8.8	hsa-mir-1908	-	83
chr11	63883159	8.11	hsa-mir-1237	+	9491
chr11	565680	4.91	hsa-mir-210	-	7482
chr11	63884358	4.64	hsa-mir-1237	+	8292
chr11	93845269	4.66	hsa-mir-548l	-	5875
chr11	558607	14.97	hsa-mir-210	-	409
chr11	558944	6.75	hsa-mir-210	-	746
chr13	90797833	13.05	hsa-mir-18a	+	3173
chr13	90797833	13.05	hsa-mir-19a	+	3313
chr13	90797833	13.05	hsa-mir-20a	+	3487
chr13	90797833	13.05	hsa-mir-19b-1	+	3614
chr13	90797833	13.05	hsa-mir-92a-1	+	3736
chr13	90798040	11.33	hsa-mir-18a	+	2966
chr13	90798040	11.33	hsa-mir-19a	+	3106
chr13	90798040	11.33	hsa-mir-20a	+	3280
chr13	90798040	11.33	hsa-mir-19b-1	+	3407
chr13	90798040	11.33	hsa-mir-92a-1	+	3529
chr13	89676530	4.8	hsa-mir-622	+	4907
chr15	94674819	5.74	hsa-mir-1469	+	2675
chr16	2258131	15.78	hsa-mir-940	+	3618
chr16	2258480	8.08	hsa-mir-940	+	3269

chr16	68157940	7.94	hsa-mir-1538	-	668
chr16	68157419	5.56	hsa-mir-1538	-	147
chr16	2257713	5.31	hsa-mir-940	+	4036
chr16	2258927	5.1	hsa-mir-940	+	2822
chr16	33872095	16.72	hsa-mir-1826	+	914
chr16	33869607	6.06	hsa-mir-1826	+	3402
chr16	33872496	5.77	hsa-mir-1826	+	513
chr17	1566694	10.46	hsa-mir-22	-	2663
chr17	54587826	8.1	hsa-mir-301a	-	4462
chr17	1905078	7.97	hsa-mir-212	-	4654
chr17	1903898	7.47	hsa-mir-212	-	3474
chr17	1905441	6.34	hsa-mir-212	-	5017
chr17	1905704	4.72	hsa-mir-212	-	5280
chr17	1907289	4.5	hsa-mir-212	-	6865
chr17	1908895	7.45	hsa-mir-212	-	8471
chr17	1908641	5.59	hsa-mir-212	-	8217
chr17	2600045	6.73	hsa-mir-1253	-	1819
chr17	30502385	98.58	hsa-mir-923	-	39
chr17	44069087	5.39	hsa-mir-196a-1	-	4167
chr17	53770484	5.34	hsa-mir-142	-	6806
chr17	1901238	5.03	hsa-mir-212	-	814
chr18	45267713	8.31	hsa-mir-1539	+	28
chr18	45267514	4.55	hsa-mir-1539	+	227
chr19	2187317	54.19	hsa-mir-1227	-	2169
chr19	2187773	15.5	hsa-mir-1227	-	2625
chr19	10375258	5.98	hsa-mir-1181	-	44
chr19	10515743	5.92	hsa-mir-1238	+	8055
chr19	3922168	5.18	hsa-mir-637	-	9658
chr19	50864675	4.83	hsa-mir-642	+	5351
chr19	13814421	6.46	hsa-mir-24-2	-	6248
chr19	13814421	6.46	hsa-mir-27a	-	6090
chr19	13814421	6.46	hsa-mir-23a	-	5948
chr19	58899173	4.89	hsa-mir-526a-1	+	2145
chr19	58899173	4.89	hsa-mir-520c	+	3346
chr19	58899173	4.89	hsa-mir-518c	+	4628
chr19	58899173	4.89	hsa-mir-524	+	6895

chr19	58899173	4.89	hsa-mir-517a	+	8161
chr19	58899173	4.89	hsa-mir-519d	+	9240
chr19	13814787	4.66	hsa-mir-24-2	-	6614
chr19	13814787	4.66	hsa-mir-27a	-	6456
chr19	13814787	4.66	hsa-mir-23a	-	6314
chr19	2191953	4.6	hsa-mir-1227	-	6805
chr2	232281283	24.81	hsa-mir-1244	+	4985
chr2	232280086	17.61	hsa-mir-1244	+	6182
chr2	232283388	8.52	hsa-mir-1244	+	2880
chr2	218972774	7.78	hsa-mir-26b	+	2839
chr2	218972515	6.73	hsa-mir-26b	+	3098
chr2	218971082	5.85	hsa-mir-26b	+	4531
chr2	70338959	5.58	hsa-mir-1285-2	-	5318
chr2	232282264	5.5	hsa-mir-1244	+	4004
chr2	232281788	4.68	hsa-mir-1244	+	4480
chr2	218973758	4.41	hsa-mir-26b	+	1855
chr2	132731092	26	hsa-mir-663b	-	-31
chr2	132736274	14.09	hsa-mir-663b	-	5151
chr2	132732348	11.04	hsa-mir-663b	-	1225
chr2	132732125	9.72	hsa-mir-663b	-	1002
chr2	132731743	6.24	hsa-mir-663b	-	620
chr2	114057804	5.07	hsa-mir-1302-3	-	661
chr2	232283751	6.89	hsa-mir-1244	+	2517
chr2	219867926	6.86	hsa-mir-153-1	-	760
chr2	176713680	6.36	hsa-mir-10b	+	9597
chr2	218969086	5.62	hsa-mir-26b	+	6527
chr20	47328481	18.61	hsa-mir-1259	+	1773
chr20	2581215	14.8	hsa-mir-1292	+	208
chr20	33506358	5.35	hsa-mir-1289-1	-	1025
chr20	62048277	6.15	hsa-mir-1914	-	4936
chr20	62048277	6.15	hsa-mir-647	-	3754
chr20	26136962	9.85	hsa-mir-663	-	48
chr20	26137237	8.93	hsa-mir-663	-	323
chr20	26138170	6.94	hsa-mir-663	-	1256
chr22	18447645	25.49	hsa-mir-1306	+	5936
chr22	18447832	10.68	hsa-mir-1306	+	5749

chr22	39817236	4.68	hsa-mir-1281	+	1227
chr22	20336362	16.14	hsa-mir-301b	+	908
chr22	20336362	16.14	hsa-mir-130b	+	1231
chr3	49041896	20.14	hsa-mir-425	-	9225
chr3	49041896	20.14	hsa-mir-191	-	8750
chr3	161600286	9.19	hsa-mir-16-2	+	4941
chr3	187983726	8.46	hsa-mir-1248	+	3429
chr3	49034541	8.29	hsa-mir-425	-	1870
chr3	49034541	8.29	hsa-mir-191	-	1395
chr3	161600034	7.38	hsa-mir-16-2	+	5193
chr3	161601500	5.1	hsa-mir-16-2	+	3727
chr3	49034247	4.53	hsa-mir-425	-	1576
chr3	49034247	4.53	hsa-mir-191	-	1101
chr4	166519681	4.63	hsa-mir-578	+	7163
chr5	54504806	12.7	hsa-mir-449b	-	2479
chr5	36188014	9.62	hsa-mir-580	-	4167
chr5	149089888	8.66	hsa-mir-378	+	2693
chr5	149091625	6.18	hsa-mir-378	+	956
chr5	179165491	6.01	hsa-mir-1229	-	7539
chr5	179166532	5.77	hsa-mir-1229	-	8580
chr5	175721268	4.79	hsa-mir-1271	+	6287
chr5	175725791	4.67	hsa-mir-1271	+	1764
chr5	88006165	7.13	hsa-mir-9-2	-	7652
chr6	32034712	10.82	hsa-mir-1236	-	2016
chr7	99537011	38.18	hsa-mir-93	-	7605
chr7	99537011	38.18	hsa-mir-25	-	7809
chr7	99537011	38.18	hsa-mir-106b	-	7378
chr7	99536438	18.23	hsa-mir-93	-	7032
chr7	99536438	18.23	hsa-mir-25	-	7236
chr7	99536438	18.23	hsa-mir-106b	-	6805
chr7	99536204	13.06	hsa-mir-93	-	6798
chr7	99536204	13.06	hsa-mir-25	-	7002
chr7	99536204	13.06	hsa-mir-106b	-	6571
chr7	30290857	8.61	hsa-mir-550-1	+	5078
chr7	30290474	5.32	hsa-mir-550-1	+	5461
chr7	30291185	4.47	hsa-mir-550-1	+	4750
chr7	25957361	8.31	hsa-mir-148a	-	1230
chr7	1034571	6.59	hsa-mir-339	-	5383

chr7	1034338	4.62	hsa-mir-339	-	5150
chr7	1033703	4.54	hsa-mir-339	-	4515
chr7	5503584	4.4	hsa-mir-589	-	1510
chr8	144970028	10.68	hsa-mir-937	-	2828
chr8	145605724	9.52	hsa-mir-1234	-	9357
chr8	145098970	6.25	hsa-mir-661	-	7535
chr8	145098389	5.11	hsa-mir-661	-	6954
chr9	95968323	17.71	hsa-let-7a-1	+	9737
chr9	95968843	10.84	hsa-let-7a-1	+	9217
chr9	95968843	10.84	hsa-let-7f-1	+	9607
chr9	130052598	4.92	hsa-mir-199b	-	5668
chrX	133510918	5.75	hsa-mir-450b	-	8960
chrX	133510918	5.75	hsa-mir-450a-1	-	8791
chrX	133510918	5.75	hsa-mir-450a-2	-	8615
chrX	133510918	5.75	hsa-mir-542	-	7785
chrX	133510918	5.75	hsa-mir-503	-	2824
chrX	133510918	5.75	hsa-mir-424	-	2511
chrX	109177247	5.98	hsa-mir-652	+	7966

References

Adachi N, Lieber MR (2002) Bidirectional gene organization: a common architectural feature of the human genome. *Cell* 109(7): 807-809

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC (2000) The genome sequence of Drosophila melanogaster. *Science* 287(5461): 2185-2195

Adhikary S, Eilers M (2005) Transcriptional regulation and transformation by Myc proteins. *Nat Rev Mol Cell Biol* 6(8): 635-645

Attwooll C, Lazzerini Denchi E, Helin K (2004) The E2F family: specific functions and overlapping interests. *EMBO J* 23(24): 4709-4716

Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefrancois P, Struhl K, Gerstein M, Snyder M (2009) Mapping accessible chromatin regions using Sono-Seq. *Proc Natl Acad Sci U S A* 106(35): 14926-14931

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, Kuznetsov H, Wang CF, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulyk ML (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324(5935): 1720-1723

Balciunaite E, Spektor A, Lents NH, Cam H, Te Riele H, Scime A, Rudnicki MA, Young R, Dynlacht BD (2005) Pocket protein complexes are recruited to distinct targets in quiescent and proliferating cells. *Mol Cell Biol* 25(18): 8166-8178

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129(4): 823-837

Bell AC, West AG, Felsenfeld G (1999) The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* 98(3): 387-396

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 125(2): 315-326

Birney E SJ, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermüller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J,

Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaöz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Löytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JN, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrímsdóttir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306(5696): 636-640

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A,

Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaoz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Loytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Xu M, Haidar JN, Yu Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrimsdottir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146): 799-816

Boon K, Caron HN, van Asperen R, Valentijn L, Hermus MC, van Sluis P, Roobeek I, Weis I, Voute PA, Schwab M, Versteeg R (2001) N-myc enhances the expression of a large set of genes functioning in ribosome biogenesis and protein synthesis. *EMBO J* 20(6): 1383-1393

Boyadjiev SA, Jabs EW (2000) Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. *Clin Genet* 57(4): 253-266

Boyle AP, Guinney J, Crawford GE, Furey TS (2008) F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* 24(21): 2537-2538

Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res* 21(3): 456-464

Burcin M, Arnold R, Lutz M, Kaiser B, Runge D, Lottspeich F, Filippova GN, Lobanenkov VV, Renkawitz R (1997) Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction with hormone receptors, is identical to the multivalent zinc finger repressor CTCF. *Mol Cell Biol* 17(3): 1281-1288

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG (2005) Natural selection on protein-coding genes in the human genome. *Nature* 437(7062): 1153-1157

Cam H, Balciunaite E, Blais A, Spektor A, Scarpulla RC, Young R, Kluger Y, Dynlacht BD (2004) A common set of gene regulatory networks links metabolism and growth inhibition. *Mol Cell* 16(3): 399-411

Campos EI, Reinberg D (2009) Histones: annotating chromatin. *Annu Rev Genet* 43: 559-599

Caretti G, Salsi V, Vecchi C, Imbriano C, Mantovani R (2003) Dynamic recruitment of NF-Y and histone acetyltransferases on cell-cycle promoters. *J Biol Chem* 278(33): 30435-30440

Chen CR, Kang Y, Siegel PM, Massague J (2002) E2F4/5 and p107 as Smad cofactors linking the TGFbeta receptor to c-myc repression. *Cell* 110(1): 19-32

Chernukhin I, Shamsuddin S, Kang SY, Bergstrom R, Kwon YW, Yu W, Whitehead J, Mukhopadhyay R, Docquier F, Farrar D, Morrison I, Vigneron M, Wu SY, Chiang CM, Loukinov D, Lobanenkov V, Ohlsson R, Klenova E (2007) CTCF interacts with and recruits the largest subunit of RNA polymerase II to CTCF target sites genome-wide. *Mol Cell Biol* 27(5): 1631-1648

Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ, Casero D, Bernal M, Huijser P, Clark AT, Kramer U, Merchant

SS, Zhang X, Jacobsen SE, Pellegrini M (2010) Relationship between nucleosome positioning and DNA methylation. *Nature* 466(7304): 388-392

Chong JL, Wenzel PL, Saenz-Robles MT, Nair V, Ferrey A, Hagan JP, Gomez YM, Sharma N, Chen HZ, Ouseph M, Wang SH, Trikha P, Culp B, Mezache L, Winton DJ, Sansom OJ, Chen D, Bremner R, Cantalupo PG, Robinson ML, Pipas JM, Leone G (2009) E2f1-3 switch from activators in progenitor cells to repressors in differentiating cells. *Nature* 462(7275): 930-934

Consortium CeS (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* 282(5396): 2012-2018

Croce CM (2009) Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet* 10(10): 704-714

Crosby ME, Almasan A (2004) Opposing roles of E2Fs in cell proliferation and death. *Cancer Biol Ther* 3(12): 1208-1211

CSH (2005) Rapid amplification of 5' complementary DNA ends (5' RACE). *Nat Methods* 2(8): 629-630

Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* 19(1): 24-32

Dai MS, Lu H (2008) Crosstalk between c-Myc and ribosome in ribosomal biogenesis and cancer. *J Cell Biochem* 105(3): 670-677

Dang CV (1999) c-Myc target genes involved in cell growth, apoptosis, and metabolism. *Mol Cell Biol* 19(1): 1-11

Dang CV, O'Donnell KA, Zeller KI, Nguyen T, Osthus RC, Li F (2006) The c-Myc target gene network. *Semin Cancer Biol* 16(4): 253-264

de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, Cramer P, Bentley D, Kornblihtt AR (2003) A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* 12(2): 525-532

De Lucia F, Dean C (2010) Long non-coding RNAs and chromatin regulation. *Curr Opin Plant Biol*

Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4(5): P3

Deschenes C, Alvarez L, Lizotte ME, Vezina A, Rivard N (2004) The nucleocytoplasmic shuttling of E2F4 is involved in the regulation of human intestinal epithelial cell proliferation and differentiation. *J Cell Physiol* 199(2): 262-273

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319): 1061-1073

Eden E, Lipson D, Yogev S, Yakhini Z (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol* 3(3): e39

Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L, Heward JM, Gough SC, de Smith A, Blakemore AI, Froguel P, Owen CJ, Pearce SH, Teixeira L, Guillevin L, Graham DS, Pusey CD, Cook HT, Vyse TJ, Aitman TJ (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat Genet* 39(6): 721-723

Farnham PJ (2009) Insights from genomic profiling of transcription factors. *Nat Rev Genet* 10(9): 605-616

Faucz FR, Horvath A, Rothenbuhler A, Almeida MQ, Libe R, Raffin-Sanson ML, Bertherat J, Carraro DM, Soares FA, Molina Gde C, Campos AH, Alexandre RB, Bendhack ML, Nesterova M, Stratakis CA (2011) Phosphodiesterase 11A (PDE11A) genetic variants may increase susceptibility to prostatic cancer. *J Clin Endocrinol Metab* 96(1): E135-140

Feagins LA, Susnow N, Zhang HY, Pearson S, Owen C, Schmalstieg WF, Terada LS, Spechler SJ, Ramirez RD, Souza RF (2006) Gain of allelic gene expression for IGF-II occurs frequently in Barrett's esophagus. *Am J Physiol Gastrointest Liver Physiol* 290(5): G871-875

Federico C, Scavo C, Cantarella CD, Motta S, Saccone S, Bernardi G (2006) Gene-rich and gene-poor chromosomal regions have different locations in the interphase nuclei of cold-blooded vertebrates. *Chromosoma* 115(2): 123-128

Filippova GN, Fagerlie S, Klenova EM, Myers C, Dehner Y, Goodwin G, Neiman PE, Collins SJ, Lobanenkov VV (1996) An exceptionally conserved transcriptional repressor,

CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol Cell Biol* 16(6): 2802-2813

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, Chew EG, Huang PY, Welboren WJ, Han Y, Ooi HS, Ariyaratne PN, Vega VB, Luo Y, Tan PY, Choy PY, Wansa KD, Zhao B, Lim KS, Leow SC, Yow JS, Joseph R, Li H, Desai KV, Thomsen JS, Lee YK, Karuturi RK, Herve T, Bourque G, Stunnenberg HG, Ruan X, Cacheux-Rataboul V, Sung WK, Liu ET, Wei CL, Cheung E, Ruan Y (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 462(7269): 58-64

Furney SJ, Higgins DG, Ouzounis CA, Lopez-Bigas N (2006) Structural and functional properties of genes involved in human cancer. *BMC Genomics* 7: 3

Gaubatz S, Lees JA, Lindeman GJ, Livingston DM (2001) E2F4 is exported from the nucleus in a CRM1-dependent manner. *Mol Cell Biol* 21(4): 1384-1392

Giangrande PH, Zhu W, Rempel RE, Laakso N, Nevins JR (2004) Combinatorial gene control involving E2F and E Box family members. *EMBO J* 23(6): 1336-1347

Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA (2004) Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* 118(5): 555-566

Gilchrist DA, Dos Santos G, Fargo DC, Xie B, Gao Y, Li L, Adelman K (2010) Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell* 143(4): 540-551

Grandori C, Cowley SM, James LP, Eisenman RN (2000) The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu Rev Cell Dev Biol* 16: 653-699

Gu J, Iyer VR (2006) PI3K signaling and miRNA expression during the response of quiescent human fibroblasts to distinct proliferative stimuli. *Genome Biol* 7(5): R42

Hansen JC (2002) Conformational dynamics of the chromatin fiber in solution: determinants, mechanisms, and functions. *Annu Rev Biophys Biomol Struct* 31: 361-392

Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM (2000) CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* 405(6785): 486-489

Hashimoto S, Suzuki Y, Kasai Y, Morohoshi K, Yamada T, Sese J, Morishita S, Sugano S, Matsushima K (2004) 5'-end SAGE for the analysis of transcriptional start sites. *Nat Biotechnol* 22(9): 1146-1149

Hatchwell E, Greally JM (2007) The potential role of epigenomic dysregulation in complex human disease. *Trends Genet* 23(11): 588-595

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459(7243): 108-112

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 39(3): 311-318

Hodges C, Bintu L, Lubkowska L, Kashlev M, Bustamante C (2009) Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science* 325(5940): 626-628

Hu Z, Killion PJ, Iyer VR (2007) Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet* 39(5): 683-687

Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, Khalil AM, Zuk O, Amit I, Rabani M, Attardi LD, Regev A, Lander ES, Jacks T, Rinn JL (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* 142(3): 409-419

Ikeda MA, Jakoi L, Nevins JR (1996) A unique role for the Rb protein in controlling E2F accumulation during cell growth and differentiation. *Proc Natl Acad Sci U S A* 93(8): 3215-3220

Initiative TAG (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408(6814): 796-815

Ip JY, Schmidt D, Pan Q, Ramani AK, Fraser AG, Odom DT, Blencowe BJ (2011) Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. *Genome Res* 21(3): 390-401

J. Craig Venter MDA, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, Robert A. Holt, Jeannine D. Gocayne, Peter Amanatides, Richard M. Ballew, Daniel H. Huson, Jennifer Russo Wortman, Qing Zhang, Chinnappa D. Kodira, Xiangqun H. Zheng, Lin Chen, Marian Skupski, Gangadharan Subramanian, Paul D. Thomas, Jinghui Zhang, George L. Gabor Miklos, Catherine Nelson, Samuel Broder, Andrew G. Clark, Joe Nadeau, Victor A. McKusick, Norton Zinder, Arnold J. Levine, Richard J. Roberts, Mel Simon, Carolyn Slayman, Michael Hunkapiller, Randall Bolanos, Arthur Delcher, Ian Dew, Daniel Fasulo, Michael Flanigan, Liliana Florea, Aaron Halpern, Sridhar Hannenhalli, Saul Kravitz, Samuel Levy, Clark Mobarry, Knut Reinert, Karin Remington, Jane Abu-Threideh, Ellen Beasley, Kendra Biddick, Vivien Bonazzi, Rhonda Brandon, Michele Cargill, Ishwar Chandramouliswaran, Rosane Charlab, Kabir Chaturvedi, Zuoming Deng, Valentina Di Francesco, Patrick Dunn, Karen Eilbeck, Carlos Evangelista, Andrei E. Gabrielian, Weiniu Gan, Wangmao Ge, Fangcheng Gong, Zhiping Gu, Ping Guan, Thomas J. Heiman, Maureen E. Higgins, Rui-Ru Ji, Zhaoxi Ke, Karen A. Ketchum, Zhongwu Lai, Yiding Lei, Zhenya Li, Jiayin Li, Yong Liang, Xiaoying Lin, Fu Lu, Gennady V. Merkulov, Natalia Milshina, Helen M. Moore, Ashwinikumar K Naik, Vaibhav A. Narayan, Beena Neelam, Deborah Nusskern, Douglas B. Rusch, Steven Salzberg, Wei Shao, Bixiong Shue, Jingtao Sun, Zhen Yuan Wang, Aihui Wang, Xin Wang, Jian Wang, Ming-Hui Wei, Ron Wides, Chunlin Xiao, Chunhua Yan, Alison Yao, Jane Ye, Ming Zhan, Weiqing Zhang, Hongyu Zhang, Qi Zhao, Liansheng Zheng, Fei Zhong, Wenyan Zhong, Shiaoping C. Zhu, Shaying Zhao, Dennis Gilbert, Suzanna Baumhueter, Gene Spier, Christine Carter, Anibal Cravchik, Trevor Woodage, Feroze Ali, Huijin An, Aderonke Awe, Danita Baldwin, Holly Baden, Mary Barnstead, Ian Barrow, Karen Beeson, Dana Busam, Amy Carver, Angela Center, Ming Lai Cheng, Liz Curry, Steve Danaher, Lionel Davenport, Raymond Desilets, Susanne Dietz, Kristina Dodson, Lisa Doup, Steven Ferriera, Neha Garg, Andres Gluecksmann, Brit Hart, Jason Haynes, Charles Haynes, Cheryl Heiner, Suzanne Hladun, Damon Hostin, Jarrett Houck, Timothy Howland, Chinyere Ibegwam, Jeffery Johnson, Francis Kalush, Lesley Kline, Shashi Koduru, Amy Love, Felecia Mann, David May, Steven McCawley, Tina McIntosh, Ivy McMullen, Mee Moy, Linda Moy, Brian Murphy, Keith Nelson, Cynthia Pfannkoch, Eric Pratts, Vinita Puri, Hina Qureshi, Matthew Reardon, Robert Rodriguez, Yu-Hui Rogers, Deanna Romblad, Bob Ruhfel, Richard Scott, Cynthia Sitter, Michelle Smallwood, Erin Stewart, Renee Strong, Ellen Suh, Reginald Thomas, Ni Ni Tint, Sukyee Tse, Claire Vech, Gary Wang, Jeremy Wetter, Sherita Williams, Monica Williams, Sandra Windsor, Emily Winn-Deen, Keriellen Wolfe, Jayshree Zaveri, Karena Zaveri, Josep F. Abril, Roderic Guigó, Michael J. Campbell, Kimmen V. Sjolander, Brian Karlak, Anish Kejariwal, Huaiyu Mi, Betty Lazareva, Thomas Hatton, Apurva Narechania, Karen Diemer, Anushya Muruganujan, Nan Guo, Shinji Sato, Vineet Bafna, Sorin Istrail, Ross Lippert, Russell Schwartz, Brian Walenz, Shibu Yooseph, David Allen, Anand Basu, James Baxendale, Louis Blick, Marcelo Caminha, John Carnes-Stine,

Parris Caulk, Yen-Hui Chiang, My Coyne, Carl Dahlke, Anne Deslattes Mays, Maria Dombroski, Michael Donnelly, Dale Ely, Shiva Esparham, Carl Fosler, Harold Gire, Stephen Glanowski, Kenneth Glasser, Anna Glodek, Mark Gorokhov, Ken Graham, Barry Gropman, Michael Harris, Jeremy Heil, Scott Henderson, Jeffrey Hoover, Donald Jennings, Catherine Jordan, James Jordan, John Kasha, Leonid Kagan, Cheryl Kraft, Alexander Levitsky, Mark Lewis, Xiangjun Liu, John Lopez, Daniel Ma, William Majoros, Joe McDaniel, Sean Murphy, Matthew Newman, Trung Nguyen, Ngoc Nguyen, Marc Nodell, Sue Pan, Jim Peck, Marshall Peterson, William Rowe, Robert Sanders, John Scott, Michael Simpson, Thomas Smith, Arlan Sprague, Timothy Stockwell, Russell Turner, Eli Venter, Mei Wang, Meiyuan Wen, David Wu, Mitchell Wu, Ashley Xia, Ali Zandieh, and Xiaohong Zhu (2001) The human genome. Science genome map. *Science* 291(5507): 1218

Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* 26(11): 1293-1300

Jimenez-Sanchez G, Childs B, Valle D (2001) Human disease genes. *Nature* 409(6822): 853-855

Jin VX, Singer GA, Agosto-Perez FJ, Liyanarachchi S, Davuluri RV (2006) Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. *BMC Bioinformatics* 7: 114

Kaestner KH, Hiemisch H, Schutz G (1998) Targeted disruption of the gene encoding hepatocyte nuclear factor 3gamma results in reduced transcription of hepatocyte-specific genes. *Mol Cell Biol* 18(7): 4245-4251

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, Taatjes DJ, Dekker J, Young RA (2010) Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 467(7314): 430-435

Kim J, Lee JH, Iyer VR (2008) Global identification of Myc target genes reveals its direct role in mitochondrial biogenesis and its E-box usage in vivo. *PLoS One* 3(3): e1798

Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* 128(6): 1231-1245

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465(7295): 182-187

Kim YH, Pollack JR (2009) Comparative genomic hybridization on spotted oligonucleotide microarrays. *Methods Mol Biol* 556: 21-32

Kimura-Yoshida C, Kitajima K, Oda-Ishii I, Tian E, Suzuki M, Yamamoto M, Suzuki T, Kobayashi M, Aizawa S, Matsuo I (2004) Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development* 131(1): 57-71

Kimura K, Wakamatsu A, Suzuki Y, Ota T, Nishikawa T, Yamashita R, Yamamoto J, Sekine M, Tsuritani K, Wakaguri H, Ishii S, Sugiyama T, Saito K, Isono Y, Irie R, Kushida N, Yoneyama T, Otsuka R, Kanda K, Yokoi T, Kondo H, Wagatsuma M, Murakawa K, Ishida S, Ishibashi T, Takahashi-Fujii A, Tanase T, Nagai K, Kikuchi H, Nakai K, Isogai T, Sugano S (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res* 16(1): 55-65

Kleinjan DA, Seawright A, Schedl A, Quinlan RA, Danes S, van Heyningen V (2001) Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. *Hum Mol Genet* 10(19): 2049-2059

Knoepfler PS, Zhang XY, Cheng PF, Gafken PR, McMahon SB, Eisenman RN (2006) Myc influences global chromatin structure. *EMBO J* 25(12): 2723-2734

Kouzarides T (2007) Chromatin modifications and their function. *Cell* 128(4): 693-705

Krumm A, Hickey LB, Groudine M (1995) Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. *Genes Dev* 9(5): 559-572

Ladomery M, Dellaire G (2002) Multifunctional zinc finger proteins in development and disease. *Ann Hum Genet* 66(Pt 5-6): 331-342

Lal A, Navarro F, Maher CA, Maliszewski LE, Yan N, O'Day E, Chowdhury D, Dykxhoorn DM, Tsai P, Hofmann O, Becker KG, Gorospe M, Hide W, Lieberman J (2009) miR-24 Inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle

genes via binding to "seedless" 3'UTR microRNA recognition elements. *Mol Cell* 35(5): 610-625

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822): 860-921

Larsen DH, Poinsignon C, Gudjonsson T, Dinant C, Payne MR, Hari FJ, Danielsen JM, Menard P, Sand JC, Stucki M, Lukas C, Bartek J, Andersen JS, Lukas J (2010) The chromatin-remodeling factor CHD4 coordinates signaling and repair after DNA damage. *J Cell Biol* 190(5): 731-740

Law SW, Conneely OM, DeMayo FJ, O'Malley BW (1992) Identification of a new brain-specific transcription factor, NURR1. *Mol Endocrinol* 6(12): 2129-2135

Lee BK, Bhinge AA, Iyer VR (2011) Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. *Nucleic Acids Res*

Lee TI, Young RA (2000) Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* 34: 77-137

Lemon B, Tjian R (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev* 14(20): 2551-2569

Levine M, Tjian R (2003) Transcription regulation and animal diversity. *Nature* 424(6945): 147-151

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18(11): 1851-1858

Li JM, Hu PP, Shen X, Yu Y, Wang XF (1997) E2F4-RB and E2F4-p107 complexes suppress gene expression by transforming growth factor beta through E2F binding sites. *Proc Natl Acad Sci U S A* 94(10): 4948-4953

Libe R, Fratticci A, Coste J, Tissier F, Horvath A, Ragazzon B, Rene-Corail F, Groussin L, Bertagna X, Raffin-Sanson ML, Stratakis CA, Bertherat J (2008) Phosphodiesterase 11A (PDE11A) and genetic predisposition to adrenocortical tumors. *Clin Cancer Res* 14(12): 4016-4024

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950): 289-293

Lin JM, Collins PJ, Trinklein ND, Fu Y, Xi H, Myers RM, Weng Z (2007) Transcription factor binding and modified histones in human bidirectional promoters. *Genome Res* 17(6): 818-827

Lindeman GJ, Gaubatz S, Livingston DM, Ginsberg D (1997) The subcellular localization of E2F-4 is cell-cycle dependent. *Proc Natl Acad Sci U S A* 94(10): 5095-5100

Liu XS (2007) Getting started in tiling microarray analysis. *PLoS Comput Biol* 3(10): 1842-1844

Logsdon CD, Fuentes MK, Huang EH, Arumugam T (2007) RAGE and RAGE ligands in cancer. *Curr Mol Med* 7(8): 777-789

Lopez-Bigas N, De S, Teichmann SA (2008) Functional protein divergence in the evolution of Homo sapiens. *Genome Biol* 9(2): R33

Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* 389(6648): 251-260

Lukas J, Petersen BO, Holm K, Bartek J, Helin K (1996) Deregulated expression of E2F family members induces S-phase entry and overcomes p16INK4A-mediated growth suppression. *Mol Cell Biol* 16(3): 1047-1057

Manuelidis L (1991) Heterochromatic features of an 11-megabase transgene in brain cells. *Proc Natl Acad Sci U S A* 88(3): 1049-1053

Marchitti SA, Orlicky DJ, Brocker C, Vasiliou V (2010) Aldehyde dehydrogenase 3B1 (ALDH3B1): immunohistochemical tissue distribution and cellular-specific localization in normal and cancerous human tissues. *J Histochem Cytochem* 58(9): 765-783

Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J, Calabrese JM, Dennis LM, Volkert TL, Gupta S, Love J, Hannett N, Sharp PA, Bartel DP, Jaenisch R, Young RA (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* 134(3): 521-533

Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 7: 29-59

McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, Keefe D, Collins FS, Willard HF, Lieb JD, Furey TS, Crawford GE, Iyer VR, Birney E (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* 328(5975): 235-239

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 28(5): 495-501

Meloni AR, Smith EJ, Nevins JR (1999) A mechanism for Rb/p130-mediated transcription repression involving recruitment of the CtBP corepressor. *Proc Natl Acad Sci U S A* 96(17): 9574-9579

Merika M, Thanos D (2001) Enhanceosomes. *Curr Opin Genet Dev* 11(2): 205-208

Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* 11(1): 31-46

Meyer N, Penn LZ (2008) Reflecting on 25 years with MYC. *Nat Rev Cancer* 8(12): 976-990

Micalizzi DS, Christensen KL, Jedlicka P, Coletta RD, Baron AE, Harrell JC, Horwitz KB, Billheimer D, Heichman KA, Welm AL, Schiemann WP, Ford HL (2009) The Six1 homeoprotein induces human mammary carcinoma cells to undergo epithelial-mesenchymal transition and metastasis in mice through increasing TGF-beta signaling. *J Clin Invest* 119(9): 2678-2690

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448(7153): 553-560

Misteli T (2007) Beyond the sequence: cellular organization of genome function. *Cell* 128(4): 787-800

Moberg K, Starz MA, Lees JA (1996) E2F-4 switches from p130 to p107 and pRB in response to cell cycle reentry. *Mol Cell Biol* 16(4): 1436-1449

Montgomery SB, Dermitzakis ET (2009) The resolution of the genetics of gene expression. *Hum Mol Genet* 18(R2): R211-215

Mosser DD, Kotzbauer PT, Sarge KD, Morimoto RI (1990) In vitro activation of heat shock transcription factor DNA-binding by calcium and biochemical conditions that affect protein conformation. *Proc Natl Acad Sci U S A* 87(10): 3748-3752

Myers LC, Kornberg RD (2000) Mediator of transcriptional regulation. *Annu Rev Biochem* 69: 729-749

Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH, Liu ET, Ruan Y (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods* 2(2): 105-111

O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT (2005) c-Myc-regulated microRNAs modulate E2F1 expression. *Nature* 435(7043): 839-843

Ohlsson R, Renkawitz R, Lobanenkov V (2001) CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet* 17(9): 520-527

Orom UA, Shiekhattar R (2011) Long non-coding RNAs and enhancers. *Curr Opin Genet Dev*

Panne D (2008) The enhanceosome. *Curr Opin Struct Biol* 18(2): 236-242

Panne D, Maniatis T, Harrison SC (2007) An atomic model of the interferon-beta enhanceosome. *Cell* 129(6): 1111-1123

Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 10(10): 669-680

Patani N, Jiang W, Mansel R, Newbold R, Mokbel K (2009) The mRNA expression of SATB1 and SATB2 in human breast cancer. *Cancer Cell Int* 9: 18

Pauli A, Rinn JL, Schier AF (2011) Non-coding RNAs as regulators of embryogenesis. *Nat Rev Genet* 12(2): 136-149

Perlmann T, Wallen-Mackenzie A (2004) Nurr1, an orphan nuclear receptor with essential functions in developing dopamine cells. *Cell Tissue Res* 318(1): 45-52

Pickering MT, Stadler BM, Kowalik TF (2009) miR-17 and miR-20a temper an E2F1-induced G1 checkpoint to regulate cell cycle progression. *Oncogene* 28(1): 140-145

Pierce AM, Gimenez-Conti IB, Schneider-Broussard R, Martinez LA, Conti CJ, Johnson DG (1998) Increased E2F1 activity induces skin tumors in mice heterozygous and nullizygous for p53. *Proc Natl Acad Sci U S A* 95(15): 8858-8863

Qian J, Esumi N, Chen Y, Wang Q, Chowers I, Zack DJ (2005) Identification of regulatory targets of tissue-specific transcription factors: application to retina-specific gene regulation. *Nucleic Acids Res* 33(11): 3479-3491

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470(7333): 279-283

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J (2010
) A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470(7333): 279-283

Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA (2010) c-Myc regulates transcriptional pause release. *Cell* 141(3): 432-445

Rando OJ, Chang HY (2009) Genome-wide views of chromatin structure. *Annu Rev Biochem* 78: 245-271

Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD (2002) E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev* 16(2): 245-256

Roman A (2006) The human papillomavirus E7 protein shines a spotlight on the pRB family member, p130. *Cell Cycle* 5(6): 567-568

Rowland BD, Bernards R (2006) Re-evaluating cell-cycle regulation by E2Fs. *Cell* 127(5): 871-874

Sampson VB, Rong NH, Han J, Yang Q, Aris V, Soteropoulos P, Petrelli NJ, Dunn SP, Krueger LJ (2007) MicroRNA let-7a down-regulates MYC and reverts MYC-induced growth in Burkitt lymphoma cells. *Cancer Res* 67(20): 9762-9770

Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* 8(6): 424-436

Schlisio S, Halperin T, Vidal M, Nevins JR (2002) Interaction of YY1 with E2Fs, mediated by RYBP, provides a mechanism for specificity of E2F function. *EMBO J* 21(21): 5775-5786

Schmidt D, Schwalie PC, Ross-Innes CS, Hurtado A, Brown GD, Carroll JS, Flicek P, Odom DT (2010) A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res* 20(5): 578-588

Schneider R, Grosschedl R (2007) Dynamics and interplay of nuclear architecture, genome organization, and gene expression. *Genes Dev* 21(23): 3027-3043

Schwartz S, Ast G (2010) Chromatin density and splicing destiny: on the cross-talk between chromatin structure and splicing. *EMBO J* 29(10): 1629-1636

Schwartz S, Meshorer E, Ast G (2009) Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol* 16(9): 990-995

Schwemmle S, Pfeifer GP (2000) Genomic structure and mutation screening of the E2F4 gene in human tumors. *Int J Cancer* 86(5): 672-677

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A* 100(26): 15776-15781

Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, Iyer VR (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol* 6(3): e65

Shivaswamy S, Iyer VR (2008) Stress-dependent dynamics of global chromatin remodeling in yeast: dual role for SWI/SNF in the heat shock stress response. *Mol Cell Biol* 28(7): 2221-2234

Sims RJ, 3rd, Mandal SS, Reinberg D (2004) Recent highlights of RNA-polymerase-II-mediated transcription. *Curr Opin Cell Biol* 16(3): 263-271

Smale ST, Kadonaga JT (2003) The RNA polymerase II core promoter. *Annu Rev Biochem* 72: 449-479

Smeenk G, Wiegant WW, Vrolijk H, Solari AP, Pastink A, van Attikum H (2010) The NuRD chromatin-remodeling complex regulates signaling and repair of DNA damage. *J Cell Biol* 190(5): 741-749

Smith AD, Sumazin P, Zhang MQ (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci U S A* 102(5): 1560-1565

Souza RF, Yin J, Smolinski KN, Zou TT, Wang S, Shi YQ, Rhyu MG, Cottrell J, Abraham JM, Biden K, Simms L, Leggett B, Bova GS, Frank T, Powell SM, Sugimura H, Young J, Harpaz N, Shimizu K, Matsubara N, Meltzer SJ (1997) Frequent mutation of the E2F-4 cell cycle gene in primary human gastrointestinal tumors. *Cancer Res* 57(12): 2350-2353

Spector DL (2004) Stopping for FISH and chips along the chromatin fiber superhighway. *Mol Cell* 15(6): 844-846

Sproul D, Gilbert N, Bickmore WA (2005) The role of chromatin structure in regulating the expression of clustered genes. *Nat Rev Genet* 6(10): 775-781

Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM (2005) Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell* 123(6): 1133-1146

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavare S, Deloukas P, Dermitzakis ET (2007) Population genomics of human gene expression. *Nat Genet* 39(10): 1217-1224

Sun H, Wu J, Wickramasinghe P, Pal S, Gupta R, Bhattacharyya A, Agosto-Perez FJ, Showe LC, Huang TH, Davuluri RV (2011) Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. *Nucleic Acids Res* 39(1): 190-201

Sun M, Hurst LD, Carmichael GG, Chen J (2005) Evidence for a preferential targeting of 3'-UTRs by cis-encoded natural antisense transcripts. *Nucleic Acids Res* 33(17): 5533-5543

Taatjes DJ (2010) The human Mediator complex: a versatile, genome-wide regulator of transcription. *Trends Biochem Sci* 35(6): 315-322

Tabach Y, Brosh R, Buganim Y, Reiner A, Zuk O, Yitzhaky A, Koudritsky M, Rotter V, Domany E (2007) Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. *PLoS One* 2(8): e807

Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM (2004) An abundance of bidirectional promoters in the human genome. *Genome Res* 14(1): 62-66

Valenzuela L, Kamakaka RT (2006) Chromatin insulators. *Annu Rev Genet* 40: 107-138

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* 5(9): 829-834

van der Sman J, Thomas NS, Lam EW (1999) Modulation of E2F complexes during G0 to S phase transition in human primary B-lymphocytes. *J Biol Chem* 274(17): 12009-12016

van Riggelen J, Yetil A, Felsher DW (2010) MYC as a regulator of ribosome biogenesis and protein synthesis. *Nat Rev Cancer* 10(4): 301-309

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 10(4): 252-263

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457(7231): 854-858

Visel A, Bristow J, Pennacchio LA (2007) Enhancer identification through comparative genomics. *Semin Cell Dev Biol* 18(1): 140-152

Vostrov AA, Quitschke WW (1997) The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation. *J Biol Chem* 272(52): 33353-33359

Vostrov AA, Taheny MJ, Quitschke WW (2002) A region to the N-terminal side of the CTCF zinc finger domain is essential for activating transcription from the amyloid precursor protein promoter. *J Biol Chem* 277(2): 1619-1627

Walsh CP, Bestor TH (1999) Cytosine methylation and mammalian development. *Genes Dev* 13(1): 26-34

Wang D, Russell JL, Johnson DG (2000) E2F4 and E2F1 have similar proliferative properties but different apoptotic and oncogenic properties in vivo. *Mol Cell Biol* 20(10): 3417-3424

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 40(7): 897-903

West AG, Gaszner M, Felsenfeld G (2002) Insulators: many functions, many mechanisms. *Genes Dev* 16(3): 271-288

Wierstra I, Alves J (2008) The c-myc promoter: still MysterY and challenge. *Adv Cancer Res* 99: 113-333

Wood AJ, Severson AF, Meyer BJ (2010) Condensin and cohesin complexity: the expanding repertoire of functions. *Nat Rev Genet* 11(6): 391-404

Woods K, Thomson JM, Hammond SM (2007) Direct regulation of an oncogenic micro-RNA cluster by E2F transcription factors. *J Biol Chem* 282(4): 2130-2134

Xu X, Bieda M, Jin VX, Rabinovich A, Oberley MJ, Green R, Farnham PJ (2007) A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. *Genome Res* 17(11): 1550-1561

Yang J, Song K, Krebs TL, Jackson MW, Danielpour D (2008) Rb/E2F4 and Smad2/3 link survivin to TGF-beta-induced apoptosis and tumor progression. *Oncogene* 27(40): 5326-5338

Yochum GS, Cleland R, McWeeney S, Goodman RH (2007) An antisense transcript induced by Wnt/beta-catenin signaling decreases E2F4. *J Biol Chem* 282(2): 871-878

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, McCarthy MI, Hattersley AT (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316(5829): 1336-1341

Zeitlinger J, Stark A, Kellis M, Hong JW, Nechaev S, Adelman K, Levine M, Young RA (2007) RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* 39(12): 1512-1516

Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, Li Z, Lee JH, Aach J, Leproust EM, Eggan K, Church GM (2009) Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods*

Zhao YP, Chen G, Feng B, Zhang TP, Ma EL, Wu YD (2007) Microarray analysis of gene expression profile of multidrug resistance in pancreatic cancer. *Chin Med J (Engl)* 120(20): 1743-1752

Zheng N, Fraenkel E, Pabo CO, Pavletich NP (1999) Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. *Genes Dev* 13(6): 666-674

Zhu W, Giangrande PH, Nevins JR (2004) E2Fs link the control of G1/S and G2/M transcription. *EMBO J* 23(23): 4615-4626

Zill OA, Scannell D, Teytelman L, Rine J (2010) Co-evolution of transcriptional silencing proteins and the DNA elements specifying their assembly. *PLoS Biol* 8(11): e1000550

Zlatanova J, Caiafa P (2009) CTCF and its protein partners: divide and rule? *J Cell Sci* 122(Pt 9): 1275-1284

Zwicker J, Lucibello FC, Wolfraim LA, Gross C, Truss M, Engeland K, Muller R (1995) Cell cycle regulation of the cyclin A, cdc25C and cdc2 genes is based on a common mechanism of transcriptional repression. *EMBO J* 14(18): 4514-4522